Modelo de aprendizaje de máquinas para identificar variables con mayor incidencia en la deserción escolar y que predicen posibles desertores de instituciones educativas en educación regular

Gabriel J. Jaramillo y Leidy J. Calderón

Especialización en Big Data e Inteligencia de Negocios

Facultad de Ingenierías y Arquitectura

Universidad Católica Luis Amigó

Medellín, Colombia

2022

Modelo de aprendizaje de máquinas para identificar variables con mayor incidencia en la deserción escolar y que predicen posibles desertores de instituciones educativas en educación regular

Gabriel J. Jaramillo y Leidy J. Calderón

Asesor: Dr. Juan Camilo Giraldo Mejía

Trabajo de grado, presentado como requisito para optar al título de

Especialista en Big Data e Inteligencia de Negocios

Facultad de Ingenierías y Arquitectura

Universidad Católica Luis Amigó

Medellín, Colombia

2022

(Dedicatoria)

A mi compañera de trabajo de grado, quien con su disciplina, constancia y compromiso me ayudó a sacar adelante este trabajo.

Gabriel Jaramillo Ciro

Agradecimientos

A nuestros maestros de la Universidad Católica Luis Amigó que nos brindaron sus conocimientos y experiencias en el transcurso de la especialización y nos ayudaron de una u otra forma para hacer posible la realización del trabajo de grado.

A nuestros compañeros y compañeras y a todas las personas que nos apoyaron y motivaron para seguir adelante y cumplir con los objetivos de este proyecto.

Tabla de Contenido

Pág.

1. Introducción 10
2. Motivación 14
3. Planteamiento del Problema 16
4. Justificación 18
5. Objetivos 19
    5.1. Objetivo General 19
    5.2. Objetivos Específicos 19
6. Marco Metodológico 20
7. Marco Referencial 24
    7.1. Marco Teórico 24
    7.2. Marco Conceptual 29
    7.3. Marco Normativo 33
    7.4. Estado del arte 34
8. Diagrama de Arquitectura del Modelo Construido 37
9. Desarrollo del Proyecto 38
    9.1. Desarrollo del Objetivo Específico 1: Realizar la Comprensión y Preparación de los Datos Suministrados por la Secretaría de Educación de Medellín. 38
        9.1.2. Fase II. Estudio y Comprensión de los Datos 46
        9.1.3. Fase III. Análisis y Preparación de los Datos 55

9.2. Desarrollo del Objetivo Específico 2: Aplicar Diferentes Técnicas Supervisadas de clasificación sobre los Datos Preparados para Generar el Modelo Óptimo 61

9.2.1. Fase IV Modelado 61

9.3. Desarrollo del Objetivo Específico 3: Evaluar el Modelo Generado para la Identificación de las Variables Predictoras que tienen Mayor Peso o Influencia en la Deserción Estudiantil. 67

9.3.1. Fase V – Evaluación..... 67

9.4. Desarrollo del Objetivo Específico 4: Presentar los Resultados Obtenidos con el Modelo Desarrollado sobre la Predicción de Posibles Desertores de acuerdo con las Variables Obtenidas que tienen Mayor Peso. 71

9.4.1. Fase VI. Despliegue 71

10. Discusión 76
11. Conclusiones 78
12. Diagrama de la Estructura del Trabajo 81
13. Referencias..... 82
14. Anexos 91

Lista de figuras

Figura 1. *Diagrama causa y efecto de la deserción escolar* 17

Figura 2. *Diagrama metodología CRISP DM* 21

Figura 3. *Modelo de operación por procesos Alcaldía de Medellín* 40

Figura 4. *Flujograma del procedimiento PR-EDUC-068* 45

Figura 5. *Listado de variables del Sistema Integrado de Matrícula (SIMAT)* 48

Figura 6. *Listado de variables del Sistema de Información para el Monitoreo, la Prevención y el Análisis de la Deserción Escolar (SIMPADE)* 49

Figura 7. *Lista de variables eliminadas* 51

Figura 8. *Listado de variables finales* 54

Figura 9. *Proceso de ETL del trabajo* 55

Figura 10. *Visualización variable dependiente - Condición del alumno año anterior* 59

Figura 11. *Visualizaciones de las variables categóricas que contiene la base de datos* 59

Figura 12. *Balanceo variable dependiente* 60

Figura 13. *Generación de Variables Dummies* 61

Figura 14. *Muestra variables escaladas* 62

Figura 15. *Creación y evaluación del modelo Árbol de decisión* 63

Figura 16. *Estructura del árbol creado* 64

Figura 17. *Matriz de confusión del modelo de Árbol de Decisión* 64

Figura 18. *Creación y evaluación del modelo de Random Forest* 65

Figura 19. *Matriz de confusión del modelo Random Forest* 65

Figura 20. *Creación y evaluación del modelo de Regresión Logística* 66

Figura 21. *Matriz de confusión del modelo Regresión Logística* 66

Figura 22. *Curva ROC del modelo* 68

Figura 23. *Variables con mayor influencia en la deserción* 70

Figura 24. *Posibles desertores clasificados por grado escolar*..... 74

Lista de tablas

Tabla 1. *Relación de las actividades a realizar con los objetivos planteados* ... **¡Error! Marcador no definido.**

Tabla 2. *Descripción PR-EDUC-068 Implementación de estrategias de acceso y permanencia* 42

Tabla 3. *Diccionario de datos* 52

Tabla 4. *Relación atributos e indicadores* 55

Tabla 5. *Clasificaciones del modelo* 70

1. Introducción

La deserción escolar, es la interrupción o desvinculación de los estudiantes del sistema educativo de manera definitiva o temporal (Ministerio de Educación Nacional, s.f.).; la tasa de deserción intra-anual (alumnos que abandonan la escuela durante el año escolar) se ha mostrado en la historia a través de indicadores; para el caso de Medellín en los últimos años, se evidencia una tendencia hacia la disminución de la deserción escolar; para el año 2018 el Distrito tuvo una tasa del 3,68% (en el caso más alto), mientras que en el año 2019 y 2021 este indicador fue de 3,51% y 3,15%, respectivamente. (Observatorio de Trayectorias Educativas, 2021). Para el año 2020, atípico por la emergencia sanitaria a causa del COVID 19, el índice de deserción no dio cuenta de la realidad de la ciudad, incluso ni del país; esto como consecuencia de las nuevas dinámicas que hicieron parte de la educación en Colombia, tales como, la virtualidad y conforme a esta las necesidades de conectividad, equipos tecnológicos e incluso dominio de sistemas de información o plataformas educativas activadas por las instituciones educativas para impartir la educación.

Así mismo, considerando que la deserción es cambiante en el tiempo, pues está determinada por múltiples factores que inciden transversalmente en los diferentes actores que en ella intervienen tales como: las dinámicas familiares, la calidad de la educación impartida, los ambientes y entornos escolares y la motivación del estudiante por culminar su proceso académico; se hace necesario investigar sobre el tema. Por tal razón y de acuerdo con el histórico de datos de matrícula del distrito de Medellín, es viable hacer un modelo predictivo mediante el uso de técnicas de Machine Learning para lograr un mejor reconocimiento de los patrones y variables más influyentes que permiten acercarse más a la realidad del fenómeno de manera temprana en las instituciones públicas de la ciudad de Medellín; adicionalmente,

este modelo permitirá a la Secretaría de Educación de Medellín optimizar la focalización de las estrategias de permanencia hacia aquellos estudiantes que tienen mayor riesgo de desertar del sistema educativo.

Son muchas las investigaciones que se han realizado en la aplicación del aprendizaje de máquinas enfocada a la predicción de posibles factores que lleven a la deserción escolar; diferentes autores como Apaza et al., (2021), Tasnim et al., (2019), Makhloga et al., (2021), Chung & Lee (2019), Coussement et al., (2020) y Viloria et al., (2019); han abordado el tema y coinciden en que el uso de técnicas de clasificación sirve para diseñar un modelo predictivo que permita reducir la deserción escolar, detección temprana e identificación de estudiantes en riesgo de desertar; en sus investigaciones también señalan que los métodos más utilizados en estos estudios son análisis bayesianos, red neuronal, máquinas de vectores de soporte, árboles de decisión, bosques aleatorios regresión logística y K- vecinos más cercanos.

Cabe señalar que como lo comentan Gallego et al., (2021), Fonseca Grandón, (2018), Gil et al., (2021) y Xie et al., (2020); la deserción escolar se encuentra ligada a las condiciones socioeconómicas, culturales e incluso demográficas, que llevan a concluir que no obstante todo lo investigado, siempre se encontrarán nuevos aspectos para intervenir en el campo de la educación el cual a pesar de estar definido como un derecho fundamental en la Constitución Política de Colombia, requiere de intervención estatal orientada a la reducción significativa de estudiantes desmotivados que abandonan el sistema educativo.

El presente trabajo está estructurado en 12 títulos, cuyo contenido se describe a continuación. En el primer título encontramos la Introducción, allí se da un contexto general de lo que es la deserción escolar, que estudios se han realizado para su prevención y la importancia de focalizar esfuerzos para contrarrestar esta problemática. El título 2, Motivación, muestra las razones que motivaron a plantear la idea de estudio y formular la propuesta para

su solución. El título 3, Planteamiento del Problema, describe la problemática objeto de estudio, además muestra un diagrama de causa y efecto para la representación de las causas potenciales de la deserción escolar. El título 4, Justificación, presenta el por qué se aborda esta problemática en este trabajo de grado, con qué fin se realiza y a través de qué se pretende ayudar a su solución. El título 5, Objetivos, plantea la solución para la problemática y los pasos específicos para lograrla. El título 6, Marco Metodológico, detalla la metodología a utilizar para el desarrollo de este trabajo y las actividades propuestas para alcanzar cada uno de los objetivos específicos. El título 7, Marco Referencial, éste a su vez contiene el Marco Teórico, el Marco Conceptual, el Marco Normativo y el Estado del arte asociado a la deserción escolar; en el Marco Teórico encontramos teorías, ideas y planteamientos desde diferentes puntos de vista sobre los principales elementos que aborda este trabajo; el Marco Conceptual, entrega definiciones de algunos términos utilizados en el desarrollo de este trabajo, para facilitar su comprensión; el Marco Normativo presenta leyes, resoluciones y documentos que se tuvieron en cuenta para el diseño, implementación y despliegue de este estudio y por último el Estado del Arte muestra la revisión realizada de otros trabajos y estudios similares relacionados con la deserción escolar. El título 8, Desarrollo del Proyecto, entrega el desarrollo de los objetivos específicos, relacionándolos a su vez a cada una de las fases de la Metodología CRISP DM escogida para la ejecución de este trabajo: el desarrollo del primer objetivo incluye las fases I, II y III de la metodología, dentro de ellas encontraremos una descripción general de los procesos de la Alcaldía de Medellín que es la Entidad donde está ubicada la problemática y en especial la Secretaría de Educación donde está el proceso específico, los propósitos del análisis de Inteligencia de Negocios (BI), los indicadores o métricas para análisis dentro del trabajo, se muestran las base de datos iniciales y finales para el estudio, generadas a través de los sistemas de información SIMAT y SIMPADE y se presentan los informes de calidad, descripción y limpieza de los datos; el desarrollo del segundo objetivo contiene la fase IV de

CRISP DM, ella muestra las técnicas a utilizar para para el cumplimiento del objetivo del trabajo, la creación de los modelos de Árbol de Decisión, Random Forest y Regresión Logística y la comparación de estos tres modelos de acuerdo a la evaluación de cada uno de ellos; el desarrollo del objetivo 3 se da en la fase V de la metodología y entrega el informe de descripción y evaluación del modelo seleccionado como óptimo para la solución del problema planteado; y por último el desarrollo del objetivo 4 se hace en la fase VI, allí se presenta el informe de resultados de posibles desertores agrupados por grados y el informe general de resultados con las variables con mayor peso sobre la deserción. El título 9, Discusión, presenta el significado de haber realizado este estudio y su importancia. El título 10, Conclusiones, muestra las conclusiones del desarrollo de este trabajo. El título 11, Referencias, muestra los diferentes datos y tipos de publicaciones utilizadas en este trabajo. Finalmente, el Título 12, Anexos, incluye el Acuerdo de Confidencialidad firmado con la Alcaldía de Medellín para el uso de los datos.

2. Motivación

A través de la participación laboral de los autores de este trabajo de grado en el Distrito de Medellín y en particular uno de ellos directamente vinculado con la Secretaría de Educación de Medellín - Subsecretaría de Planeación Educativa; surge la necesidad de realizar un mejor acompañamiento en el ejercicio del análisis y prevención de la deserción escolar, donde se puedan identificar los factores de riesgo asociados a esta problemática y contribuir en que las estrategias de permanencia estudiantil establecidas en la Entidad respondan efectivamente a la disminución de este fenómeno; adicionalmente, la actual administración se ha comprometido en lograr una deserción total oficial del 2,5 % para el año 2023, tal y como quedó plasmado en el Plan de Desarrollo Municipal vigente (Alcaldía de Medellín, 2020); estos factores motivaron a la realización de este trabajo.

Actualmente el Distrito por medio del Observatorio para la Calidad Educativa de Medellín (OCEM), recolecta, produce, procesa y divulga información que permite y promueve el diseño y direccionamiento de estrategias que permitan transformar el sistema educativo de Medellín. Además, desde el área de Permanencia y del programa Entorno Escolar Protector de la Secretaría de Educación, se cuenta con diferentes programas enfocados a la permanencia de los estudiantes en las instituciones educativas, algunos programas son: transporte escolar, kits escolares, Programa de Alimentación Escolar (PAE), jornada escolar complementaria, Computadores Futuro, vestuario, entre otros, estos programas benefician alrededor de 260.000 estudiantes (Alcaldía de Medellín, 2022). Así mismo, se cuenta con un Sistema de Alertas Tempranas de Medellín (SATMED) que permite la identificación anticipada de alertas con el fin de evitar vulneración de los derechos humanos.

Además, desde el OCEM se vienen realizando diferentes estudios sobre las realidades del sistema educativo de Medellín, entre ellos se encuentra el estudio realizado en el año 2020 cuando emergió la pandemia producto del COVID-19; el cual cuenta las acciones más relevantes realizadas por la alcaldía de Medellín para garantizar la permanencia y la calidad de la educación durante el momento más crítico de la COVID 19 (Secretaría de Educación, 2022).

3. Planteamiento del Problema

Existen diversos factores que inciden en la deserción estudiantil; de acuerdo con Gallego et al., (2021) “El problema de la deserción escolar es multifactorial: características psicológicas de los estudiantes, familiares y condiciones sociales, organizacionales e institucionales” (p.7)., son factores que inciden fuertemente en la deserción.

La familia es un factor con un gran peso sobre la deserción; situaciones como agresiones físicas y verbales de los padres y falta de apoyo pueden desencadenar comportamientos desfavorables en el desempeño estudiantil; así mismo las dificultades económicas del hogar afectan la calidad de vida de los estudiantes y el acceso a la educación.

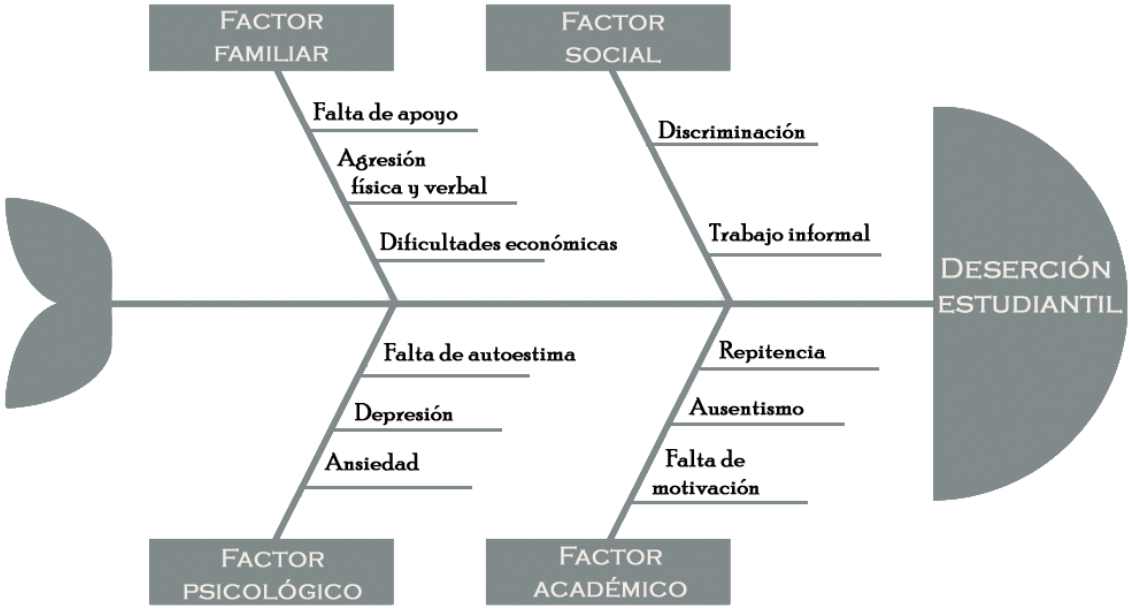
Un segundo factor que genera impedimentos para la continuidad de los alumnos en su vida estudiantil es el factor social; dentro de él encontramos el trabajo informal que desarrolla el estudiante por su condición económica y que le impide dedicar el tiempo suficiente al estudio; otra causa son las discriminaciones y tratos de manera desigual que sufre el estudiante en su entorno social, éstos traen consecuencias en los diferentes roles en los que se desenvuelve el estudiante. En general, la falta de apoyo tanto familiar como social son factores fundamentales para la deserción de los estudiantes, debido a que estos son el contexto del compromiso y socialización del niño y del adolescente en la escuela (Gil et al., 2021).

El factor académico también juega un papel importante para la permanencia de los estudiantes; contenidos académicos apropiados y buenas relaciones maestro-alumno, ayudan a combatir causas del abandono escolar como el ausentismo y la falta de motivación; otra causa a nivel académico que también inciden en el abandono es la dificultad de adaptación nuevamente de los estudiantes que son repitentes.

Un cuarto factor y quizá uno de los más relevantes que intervienen en el abandono escolar, es el psicológico; la falta de autoestima, la depresión y la ansiedad, entre otros, son causas a nivel psicológico que generan que los estudiantes no se sientan en capacidad de continuar con su proceso formativo y decidan desertar.

Estos factores y muchos otros que puedan surgir, son los que se deben combatir para lograr la reducción de la deserción escolar; la motivación en todos los ámbitos es un punto clave para lograrlo, dado que, cuando el estudiante está expuesto a múltiples factores de riesgo, es más probable que haya menos motivación para hacer el trabajo escolar y esto conlleve el abandono escolar (Gil et al., 2021).

Figura 1. *Diagrama causa y efecto de la deserción escolar*



4. Justificación

Considerando que la deserción escolar es una problemática que trae riesgos psicosociales, aumento en la tasa de desempleo, trabajo informal, entre otros; desde la Secretaría de Educación, se requiere disminuir la tasa de deserción estudiantil y mejorar los procesos de reinversión del recurso en estrategias de permanencia, en el ejercicio de garantizar el derecho fundamental a la educación para así eliminar las barreras de la ignorancia, del desconocimiento y del atraso en la construcción de una mejor calidad de vida. Por tal razón, es necesario hacer un modelo predictivo mediante el uso de técnicas de Machine Learning para que esta problemática se mitigue de manera temprana en las instituciones públicas de la ciudad de Medellín. Además, con este modelo también se busca optimizar la focalización de las estrategias de permanencia hacia aquellos estudiantes que tienen mayor riesgo de desertar del sistema educativo.

Con el uso e implementación de técnicas de Machine Learning en la solución de la problemática planteada en este trabajo, se logra un factor diferencial con relación a otras soluciones planteadas en la Secretaría de Educación, dado que se puede realizar automatización e identificación de patrones con mínima intervención humana, lo que llevaría a optimizar recursos y tiempos en relación con las estrategias que vienen utilizando hoy en día.

5. Objetivos

5.1. Objetivo General

Construir un modelo de aprendizaje de máquinas para identificar las variables con mayor incidencia en la deserción escolar y que predicen posibles desertores de instituciones educativas en educación regular.

5.2. Objetivos Específicos

- Realizar la comprensión y preparación de los datos suministrados por la Secretaría de Educación de Medellín.
- Aplicar diferentes técnicas supervisadas de clasificación sobre los datos preparados para generar el modelo óptimo.
- Evaluar el modelo generado para la identificación de las variables predictoras que tienen mayor peso o influencia en la deserción estudiantil.
- Presentar los resultados obtenidos con el modelo desarrollado sobre la predicción de posibles desertores de acuerdo con las variables obtenidas que tienen mayor peso.

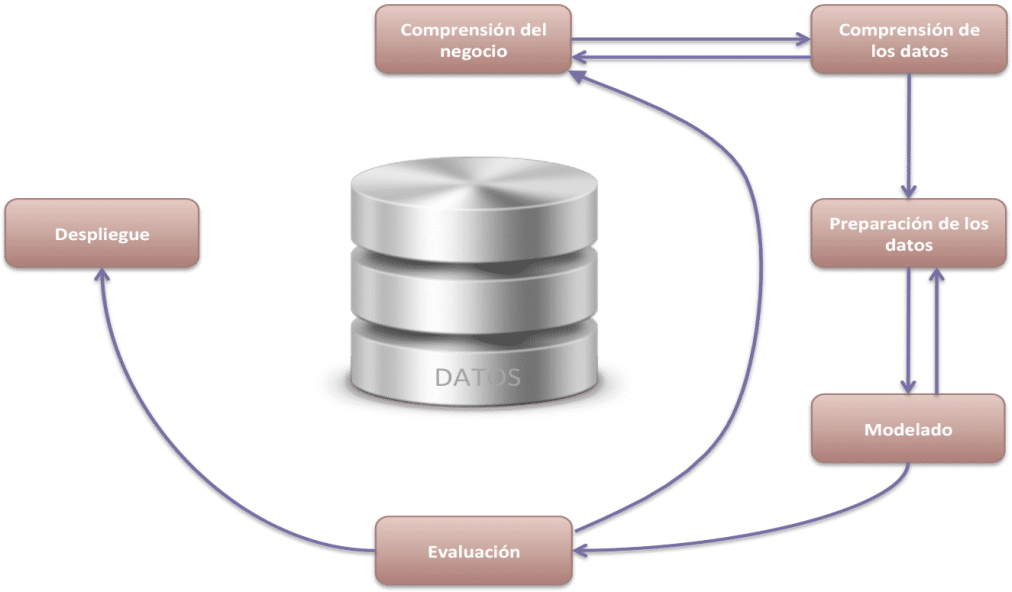
6. Marco Metodológico

Para alcanzar los objetivos propuestos en este trabajo, se utilizará la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) con sus 6 fases; esta metodología integra todas las actividades necesarias para el desarrollo de este estudio, desde la fase inicial de comprensión del negocio hasta el despliegue del modelo predictivo propuesto como solución a través del uso de técnicas de Machine Learning.

Tabla 1. *Relación de las actividades a realizar con los objetivos planteados*

Objetivo específico	Actividad	Entregable
1. Realizar la comprensión y preparación de los datos suministrados por la Secretaría de Educación de Medellín	Examinar la situación actual de los procesos de la Secretaría de Educación en los que se desarrollan las estrategias para la prevención de deserción escolar.	Informe de entendimiento del negocio
	Conseguir las bases de datos con las variables necesarias determinadas.	Bases de datos de estudiantes de la ciudad de Medellín.
		Autorización uso de datos
	Identificar los problemas que las bases tengan en cuanto a la calidad de los datos.	Informe de calidad de datos
	Generar informe con la descripción de los datos.	Informe de descripción de los datos
	Realizar la limpieza de los datos seleccionados.	Informe con la limpieza de datos realizada
	Integrar los datos seleccionados como necesarios de las bases conseguidas desde los diferentes sistemas de información.	Base con datos integrados para realizar el estudio
2. Aplicar diferentes técnicas supervisadas de clasificación sobre los datos preparados para generar el modelo óptimo.	Seleccionar las técnicas de modelado a aplicar, acorde con los datos y objetivos.	Lista de técnicas a utilizar
	Construir los modelos a partir de la aplicación de las técnicas seleccionadas.	Informe de construcción de los modelos
3. Evaluar el modelo generado para la identificación de las variables predictoras que tienen mayor peso o influencia en la deserción estudiantil.	Utilizar el conjunto de datos de test sobre los modelos generados para su evaluación.	Informe de evaluación de los modelos
	Realizar ajustes en el proceso, en caso de ser necesario	Nuevas actividades a realizar
	Realizar revisión de los resultados obtenidos con el modelo construido respecto con los propósitos de analítica propuestos.	Informe de revisión resultados vs objetivos
4. Presentar los resultados obtenidos con el modelo desarrollado sobre la predicción de posibles desertores de acuerdo con las variables obtenidas que tienen mayor peso.	Aplicar el modelo construido en una base de datos de matrícula vigente de instituciones educativas de la ciudad de Medellín	Informe general de resultados del proyecto y de la aplicación del modelo
	Generar un informe con los resultados obtenidos de acuerdo con los objetivos del proyecto.	

Figura 2. *Diagrama metodología CRISP DM*



Nota: Tomado de *Metodología CRISP-DM*, por Daniel Álvarez Gil, 2021, Adictos al Trabajo.

Fase I. Entendimiento o comprensión del negocio

A partir del uso de bases de datos de matrícula de las Instituciones Educativas de la ciudad Medellín, se pretende realizar un modelo por medio de técnicas supervisadas que permita detectar posibles desertores; para esto se llevará a cabo una actividad derivada del primer objetivo de esta investigación; la cual consiste en examinar la situación actual de los procesos de la Secretaría de Educación en los que se desarrollan las estrategias para la prevención de deserción escolar.

Fase II. Estudio y comprensión de los datos

Las bases que se van a emplear cuentan con datos del estudiante, de su núcleo familiar, del contexto institucional y municipal; lo que genera información para el monitoreo, la prevención y el análisis de la deserción escolar; de igual manera esta fase responde al primer objetivo e incluye las siguientes tres actividades para su desarrollo; conseguir las bases de

datos con las variables necesarias para la predicción de la deserción escolar, identificar los problemas que las bases tengan en cuanto a la calidad de los datos y generar informe con la descripción de los datos.

Fase III. Preparación de los datos

En esta última fase del primer objetivo, se desarrollarán dos actividades en aras de lograr una buena calidad de datos; para ello, primero se realiza la limpieza de los datos seleccionados y al final se integran los datos seleccionados como necesarios, de las bases adquiridas desde los diferentes sistemas de información.

Fase IV. Modelado

En esta fase se implementarán técnicas supervisadas de clasificación de Machine Learning en busca de la construcción de un modelo que permita alcanzar los objetivos del proyecto, entregando los pesos de las variables dadas para identificar cuáles son los atributos que más influyen en la deserción escolar. Esta fase da respuesta al segundo objetivo de la investigación y se desarrolla a partir de dos actividades: seleccionar las técnicas de modelado a aplicar, acorde con nuestros datos y objetivos y construir los modelos a partir de la aplicación de las técnicas seleccionadas, definiendo los conjuntos de datos de entrenamiento y test y ajustando los parámetros utilizados en los modelos de acuerdo con los objetivos.

Fase V. Evaluación (obtención de resultados)

Esta fase va en línea con el tercer objetivo de este trabajo, el cual incluye tres actividades; aplicar el conjunto de datos de test sobre los modelos generados para su evaluación, realizar los ajustes en el proceso, en caso de ser necesario y realizar la revisión de

los resultados de evaluación obtenidos con el modelo construido versus los propósitos de analítica propuestos.

Fase VI. Despliegue (puesta en producción)

Para el último objetivo de este trabajo desarrollamos a través de dos actividades la fase de despliegue, en ella se procederá a aplicar el modelo construido identificado como óptimo en una base de datos de matrícula vigente de instituciones educativas de la ciudad de Medellín, del sector oficial y de matrícula en edad regular (grado de 0 a 11) y como cierre del trabajo, se generará un informe con los resultados obtenidos de acuerdo con los objetivos del proyecto.

7. Marco Referencial

7.1. Marco Teórico

En Colombia “la educación es un derecho de la persona y un servicio público que tiene una función social” (Constitución Política de Colombia 1991. Art. 67); mediante la cual se forjan personas que con el tiempo construyen una vida con calidad, a través de las oportunidades a las que pueden acceder a partir de las habilidades y los conocimientos adquiridos. En Colombia el Sistema Educativo está conformado por 6 etapas: la Educación Inicial, Preescolar, Básica (primaria cinco grados y secundaria cuatro grados), la Educación Media (dos grados y culmina con el título de bachiller), Superior y la Educación para el Trabajo y el Talento Humano. (Ministerio de Educación, s.f.). En general la educación entrega las herramientas necesarias para una formación integral; sin embargo, este proceso va de la mano de la familia, como principal autoridad en la educación de los niños, además del Estado y la sociedad quienes “son responsables de la educación, que será obligatoria entre los cinco y los quince años de edad y que comprenderá como mínimo, un año de preescolar y nueve de educación básica” (Constitución Política de Colombia 1991. Art. 67); desafortunadamente, no siempre las condiciones en las que el estudiante desarrolla su vida estudiantil permiten cumplir con esta obligatoriedad, lo que trunca el proceso educativo y lleva a que una persona abandone sus estudios.

Existen muchas teorías y modelos en materia de retención escolar, la mayoría de ellos se centran en el equilibrio académico y social; Vincent Tinto aporta la teoría más citada por diferentes autores; Tinto (1975), menciona que “una persona con una insuficiente integración en la estructura social tiene más probabilidad de suicidarse… las condiciones que llevan a una persona al suicidio, son semejantes a aquéllas que conducen al estudiante al abandono

escolar" (p. 2-3).; sin embargo, indica que no son suficientes para explicar el abandono, también hace referencia a factores relacionados con la persistencia educativa, condición social, antecedentes académicos, aptitudes, expectativas y motivaciones del estudiante (Tinto, 1975, p. 3); en 1989 Tinto amplía su teoría y muestra cinco grupos para asociar los factores que influyen en el abandono, los psicológicos, sociales, económicos, organizacionales e interacciónales (p.3). Diaz Peralta (2008), propone un modelo conceptual de deserción estudiantil para explicar el fenómeno de la deserción en las universidades chilenas.; en él se representa en forma gráfica al estudiante, su dinámica con los factores que intervienen en la deserción y su condición de equilibrio y cambio. El modelo asume que todos los factores actúan en forma permanente sobre el estudiante durante sus años de estudios; sin embargo, el nivel motivacional va cambiando con los años, pues se relaciona directamente con la integración académica y la integración social, cuando se rompe este equilibrio, el estudiante abandona. Escudero (2015), habla del abandono como el fracaso escolar y argumenta que está ligado a la escuela como una institución que tiene sus propias reglas de juego para formar a los estudiantes en un determinado sistema de valores, conocimientos, capacidades y formas de vida; por lo tanto, no es un fenómeno natural, sino una realidad construida en y por la escuela en sus relaciones con los estudiantes y de éstos con ella (p. 2).

Como se ha mostrado; aunque es vista desde diferentes enfoques; la deserción escolar es la consecuencia de la integración de diferentes variables que no están equilibradas o que se van desequilibrando durante la vida escolar, es un proceso cíclico que se va ajustando de acuerdo con los cambios de los diferentes entornos en los que se desenvuelve el estudiante; tradicionalmente, el Ministerio de Educación Nacional Colombiano ha medido la deserción a través de la tasa de deserción intra-anual, es decir, el porcentaje de estudiantes que dejan de estudiar durante el transcurso del año académico, en comparación con los inicialmente matriculados. (Ministerio de Educación Nacional (p. 2). A nivel mundial se han desarrollado

diferentes estudios para combatir la problemática de la deserción escolar, muchos de ellos basados en Machine Learning; Tom Mitchell (1997), en su libro define Machine Learning como el estudio de algoritmos informáticos que mejoran automáticamente a través de la experiencia; ahora veamos cómo aparece el Machine Learning: Alan Turing matemático inglés fue pionero en investigación en computación y "Machine Learning" durante los años 1940 y 1950. Turing introdujo el Test de Turning; nombrado así en su honor, en su trabajo de 1950 llamado "Computing Machinery and Intelligence", propuso que para decir que un ordenador posee verdadera inteligencia, este debe ser capaz de imitar las respuestas que daría un humano ante condiciones específicas. El test de Turing original requería de tres terminales separadas entre sí. Una terminal sería operada por un ordenador, mientras que las otras dos serían operadas por humanos. Durante la prueba, uno de los humanos tiene la tarea de hacer preguntas, mientras que el otro humano y el ordenador deberán responderlas. (Rubio, 2022)

El profesor Arthur L. Samuel, pionero de la investigación de inteligencia artificial; en 1952, escribe parece ser, el primer programa de autoaprendizaje del mundo capaz de aprender. El software era un programa que jugaba a las damas y que mejoraba su juego partida tras partida. En el año 1962, los primeros programas tenían todavía algunos defectos bastante graves debido a la linealidad del polinomio de evaluación. Samuel perseveró en una larga serie de experimentos y modificaciones del programa. Cinco años más tarde publicó el segundo artículo sobre aprendizaje automático, en el que describe algunas mejoras radicales y técnicas recientemente desarrolladas, incluido su famoso algoritmo de tabla de firmas para combinar parámetros de forma no lineal. (Wiederhold, G., McCarthy, J. 1992)

En 1981 Gerald Dejong introduce el concepto de aprendizaje basado en explicaciones (EBL), en el que una computadora analiza los datos de entrenamiento y crea una regla general que puede seguir descartando datos sin importancia; ya para los años 90's el trabajo en

Machine Learning gira desde un enfoque orientado al conocimiento (knowledge-driven) hacia uno orientado al dato (data-driven). Los científicos comienzan a crear programas que analizan grandes cantidades de datos y extraen conclusiones de los resultados. (Marr, 2016)

La importancia de los datos es fundamentalmente central en el proceso de aprendizaje automático; actualmente el Machine Learning se concibe como el funcionamiento calibrado de algoritmos y modelos; los algoritmos de aprendizaje automático son metodologías matemáticas que producen un conjunto de resultados con la ayuda de los datos estructurados o no estructurados proporcionados. La eficiencia de un programa de Inteligencia Artificial impulsado por Machine Learning depende de la calidad de los datos de entrenamiento que se introducen en el código del algoritmo. Los conjuntos de datos inexactos también pueden degradar el rendimiento. El aprendizaje automático se usa para manejar tareas computacionales complejas que involucran enormes cantidades de datos y ninguna fórmula estática para obtener el resultado. A lo largo de los años, a medida que el estudio y la evolución han continuado en el aprendizaje automático, sectores empresariales como el médico, la producción de energía, la automoción, la industria aeroespacial, la fabricación y las finanzas se han beneficiado de sus modelos. Los modelos y algoritmos de aprendizaje automático están ayudando a resolver problemas específicos del sector y brindan soluciones futuristas para toda la industria mediante la detección de objetos, la calificación crediticia, la previsión comercial, la secuenciación de ADN y el mantenimiento predictivo. (La aplicación de algoritmos basados en datos en el aprendizaje automático, 2020)

El aprendizaje automático tiene algunos modelos, entre ellos se encuentra el aprendizaje supervisado de clasificación o también conocido como métodos de clasificación, el cual se basa en tener un conjunto de datos de entrenamiento etiquetados los cuales contienen muestras normales como anómalas para construir un modelo predictivo (Omar et al, 2013).

También (Cunningham & Delany, 2008) en su libro Técnicas de aprendizaje automático para multimedia indican que el aprendizaje supervisado implica aprender un mapeo entre conjuntos de variables de entrada y de salida, donde las variables de salida son las etiquetas para predecir los datos deseados. Además, en los modelos de clasificación las variables dependientes (salida) corresponden a atributos que indican a qué clase en particular pertenece una muestra (González, 2015). Los algoritmos supervisados de clasificación más comunes son, redes neuronales (NN), máquinas de vectores de soporte (SVM), k-vecinos más cercanos (k-NN), redes bayesianas (BN), bosques aleatorios (RF) y árbol de decisión (DT).

Los algoritmos de redes neuronales son nodos interconectados, los cuales están organizados por capas. Las redes neuronales pueden aprender de forma supervisada, por lo cual intenta predecir los resultados comparando sus predicciones con la respuesta objetivo y aprende de sus errores (Santamaría, 2015). Las máquinas de vectores de soporte según abril, 2003, “son sistemas de aprendizaje que utilizan como espacio de hipótesis, funciones lineales en espacios característicos de dimensión muy alta, ensayando algoritmos de aprendizaje de la teoría de la optimización que implementan un aprendizaje sesgado derivado a partir de la teoría del aprendizaje estadístico”. Los árboles de decisión se crean a partir de diagramas de flujos. Estos pueden clasificar grupos de variables o valores de variables dependientes en función de variables independientes. Algunas de las ventajas de utilizar un árbol de decisión son: Facilita la interpretación de la decisión adoptada, facilita la comprensión del conocimiento utilizado en la toma de decisiones, explica el comportamiento respecto a una determinada decisión y reduce el número de variables independientes (Berlanga et al, 2013).

Los métodos de clasificación de bosques aleatorios consisten en una combinación de clasificadores de árboles, donde cada árbol genera un voto individual para la clase más popular con el fin de clasificar un vector de entrada. Una de las principales ventajas de los bosques

aleatorios, es que cada vez que un árbol crece en profundidad y genera nuevos datos de entrenamiento usando combinaciones de funciones, estos no se podan. Según estudios la elección de métodos de poda y no medidas de selección de atributos, estos afectan el rendimiento de los clasificadores basados en árboles (Pal, 2005).

7.2. Marco Conceptual

Deserción Escolar:

Puede entenderse como el abandono del sistema escolar por parte de los estudiantes, provocado por la combinación de factores que se generan tanto al interior del sistema como en contextos de tipo social, familiar, individual y del entorno. La tasa de deserción intra-anual solo tiene en cuenta a los alumnos que abandonan la escuela durante el año escolar, ésta se complementa con la tasa de deserción interanual que calcula aquellos que desertan al terminar el año escolar. (Ministerio de Educación Nacional, 2022a)

Educación Básica:

“Con una duración de nueve grados escolares, comprendidos en dos ciclos: la educación básica primaria de cinco grados y la educación básica secundaria de cuatro grados” (Ministerio de Educación Nacional, 2022b).

Educación Media:

“Se compone de dos grados escolares 10 y 11 y culmina con el título de bachiller” (Ministerio de Educación Nacional, 2022b).

Permanencia Estudiantil:

“lapso de tiempo que toma, desde el ingreso hasta obtener su título profesional” (Murillo-Zabala & Jurado-De los Santos, 2020, p. 20).

Ausentismo:

"El ausentismo escolar se define habitualmente como la inasistencia reiterada o prolongada a clases durante el año escolar de un estudiante" (Pavez, 2020, p. 3).

Educación Temprana:

"la educación o estimulación temprana es un conjunto de técnicas de intervención educativas que pretende impulsar el desarrollo cognitivo, social y emocional del niño durante la etapa infantil (de 0 a 6 años)" (la Universidad en Internet (UNIR), 2020).

Control Cognitivo:

"es descrito como la habilidad para inhibir una respuesta preponderante y con cierto grado de automaticidad, en favor de otras respuestas que necesitan de la puesta en marcha de procesos atencionales más elaborados" (Gutiérrez-Cobo et al., 2017, p. 6).

Machine Learning:

Es una disciplina del campo de la Inteligencia Artificial utilizada en este trabajo para identificar a través de algoritmos, patrones en datos masivos y elaborar las predicciones propuestas para la problemática planteada. Este aprendizaje permite a los computadores realizar tareas específicas de forma autónoma, es decir, sin necesidad de ser programados. (Iberdrola, s.f.)

Big Data:

Conjunto de tecnologías diseñadas para almacenar, analizar y gestionar grandes volúmenes de datos, como los datos recolectados para este trabajo y que permitirá desarrollar soluciones inteligentes. (Iberdrola, s.f.)

Minería de datos:

Permitirá estudiar métodos y algoritmos para la extracción automática de información sintetizada sobre deserción escolar que permita caracterizar las relaciones escondidas en la gran cantidad de datos que se tienen (Beltrán Beatriz, s.f.).

Aprendizaje supervisado:

Estos algoritmos cuentan con un aprendizaje previo basado en un sistema de etiquetas asociadas a unos datos que les permiten tomar decisiones o hacer predicciones como la que se busca en este trabajo sobre la deserción escolar. (Iberdrola s.f.)

Regresión lineal:

Técnica que puede ser utilizada para el desarrollo de los objetivos de este trabajo; se implementa identificando una variable dependiente (y) y todas las variables independientes (X1, X2.). Se asume que la relación entre estas y aquella es lineal. Todas las variables han de ser continuas. El resultado es la ecuación de la recta que mejor se ajusta al juego de datos y esta ecuación se interpreta o se usa para predicción. (Beltrán Beatriz, s.f.)

Bosques aleatorios:

Es un algoritmo de aprendizaje supervisado que puede ser utilizado para el desarrollo de los objetivos de este trabajo; como ya se puede ver en su nombre, crea un bosque y lo hace de alguna manera aleatorio. Para decirlo en palabras simples: el Bosque Aleatorio crea múltiples árboles de decisión y los combina para obtener una predicción más precisa y estable. (AprendeIA, 2019)

Árboles de decisión:

Son herramientas analíticas que pueden ser utilizadas para el desarrollo de los objetivos de este trabajo; las cuales se emplean para el descubrimiento de reglas y relaciones mediante la ruptura y subdivisión sistemática de la información contenida en el conjunto de datos" (Beltrán Beatriz, s.f.).

Redes Bayesianas:

Técnica que puede ser utilizada para el desarrollo de los objetivos de este trabajo; la cual tiene varias ventajas: permiten aprender sobre relaciones de dependencia y causalidad, permiten

combinar conocimiento con datos, evitan el sobre-ajuste de los datos y pueden manejar bases de datos incompletas. (Beltrán Beatriz, s.f.)

Redes Neuronales:

Técnica que puede ser utilizada para el desarrollo de los objetivos de este trabajo; son capaces de detectar y aprender patrones y características dentro de los datos" (Beltrán Beatriz, s.f.).

Regresión logística:

Es un algoritmo de clasificación que puede ser utilizado para el desarrollo de los objetivos de este trabajo; se utiliza para predecir la probabilidad de una variable dependiente categórica. En la regresión logística, la variable dependiente es una variable binaria que contiene datos codificados como 1 – 0, sí – no, abierto – cerrado, etc. (AprendeIA, 2019)

Metodología CRISP DM:

Esta metodología integra todas las tareas necesarias para el desarrollo de este estudio, desde la fase inicial de comprensión del problema hasta el despliegue del modelo predictivo propuesto como solución a través del uso de técnicas de Machine Learning, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. (IBM, 2021)

Google Colab:

Es un documento ejecutable que permitirá escribir, ejecutar y compartir código arbitrario de Python en el navegador, para realizar las tareas de aprendizaje automático y análisis de los datos disponibles para este trabajo. (Agrega, s.f.)

Python:

Lenguaje de programación utilizado para el desarrollo de este trabajo, ampliamente utilizado en la ciencia de datos y Machine Learning (ML). Es eficiente y fácil de aprender, además se puede ejecutar en muchas plataformas diferentes. (AWS, s.f.)

7.3. Marco Normativo

- Resolución 07797 de 2015 Por medio de la cual se establece el proceso de gestión de la cobertura educativa en las Entidades Territoriales Certificadas.

Capítulo ll. De los responsables y sus competencias generales en el marco del proceso de gestión de la cobertura educativa en sus artículos del 4 al 8.

Capítulo V. Reportes de información y cronograma del proceso de gestión de la cobertura educativa en sus artículos del 29 al 31.

- Ley 1581 de 2012 Protección de Datos Personales.

Firma del Acuerdo de Confidencialidad para utilizar las bases de datos de los estudiantes de la ciudad, con la Alcaldía de Medellín.

- Plan de Desarrollo Medellín Futuro 2020 - 2023

En su numeral 3.2.3. Componente. Educación para todos y todas, tiene como objetivo garantizar la cobertura universal del derecho a la educación de calidad en la primera infancia, la educación básica y la media técnica en condiciones de acceso, permanencia, inclusión, equidad, igualdad, diversidad y ambientes de aprendizaje; y ampliar las oportunidades de acceso a la educación postsecundaria para una Medellín Futuro digna y en paz; establece como meta para 2020- 2023 un indicador de tasa de deserción total en edad escolar del sector oficial del 2.5%.

- Procedimiento interno Alcaldía de Medellín PR-EDUC-068 Implementación de estrategias de acceso y permanencia.

Mediante el cual se realiza seguimiento a las estrategias actuales y se identifican e implementan estrategias de acceso y permanencia pertinentes, que permitan el ingreso y la continuidad en el servicio educativo a los niños, adolescentes, jóvenes y adultos de la ciudad de Medellín.

7.4. Estado del arte

El Big Data y el Machine Learning a través del uso de diversos métodos está aportando a la predicción de la deserción estudiantil, lo que puede permitir establecer estrategias para frenar este fenómeno en el sector educativo. De acuerdo con Apaza et al., (2021)

> La predicción ayudará a proyectar estrategias que en conjunto con la institución, docentes, estudiantes y padres de familia puedan mejorar sus actividades del proceso de enseñanza-aprendizaje. Para lograr el propósito de la predicción, se utilizará Machine Learning, en concreto, técnicas de clasificación para diseñar un modelo predictivo que permita determinar el rendimiento académico de los alumnos y reducir su deserción, así como determinar el mejor algoritmo predictivo. (p.1)

Para reducir la deserción escolar, primero es necesario identificar las razones detrás de la deserción del estudiante, para ello la minería de datos es el método con el cual se puede aprender información interesante y útil a partir de una gran cantidad de datos. Este método comprende la naturaleza predictiva (aprendizaje supervisado) y descriptiva (aprendizaje no supervisado) del aprendizaje. Algunos de los métodos utilizados para identificar los estudiantes que abandonan son los árboles de decisiones, algoritmos de clasificación (Naive Bayes, Neural Network, Support Vector Machine, Decision Tree y Random Forest) (Tasnim et al., 2019). Igualmente, según Maheshwari et al., (2020) empleando el uso de la minería de datos educativos por medio de una metodología y un algoritmo de agrupamiento específico para identificar los factores que provocan la deserción de los estudiantes en los diferentes niveles educativos, como primaria, secundaria y bachillerato, y también su porcentaje de impacto entre los estudiantes, dará una solución a este problema. Asimismo, se podrá predecir con el uso de la minería de datos educativos por medio del aprendizaje automático con métodos como la regresión logística, árboles de decisión y K-vecinos más cercanos; si un estudiante abandonará

o continuará su educación, lo que podrá permitir al profesor proporcionar un monitoreo y asesoramiento necesario para el alumno (Makhloga et al., 2021). Por otro lado, otras técnicas predictivas de aprendizaje que se utilizan para la predicción temprana de la deserción de los estudiantes son el modelo bidireccional a largo y corto plazo, el cual tiene diversas características para evaluar cómo se podría desempeñar un estudiante nuevo y el método de campo aleatorio de condiciones, el cual es un etiquetado de secuencias que permite identificar la etiqueta de cada estudiante de forma independiente (Uliyan et al., 2021).

De acuerdo con las investigaciones realizadas, se han implementado diversos modelos de Big Data/Machine Learning para la predicción de la deserción escolar, algunos casos son como el de Corea, donde por medio de modelado predictivo usando aprendizaje automático con Big Data se utilizó el modelo de bosques aleatorios con datos sobre la asistencia y actividades de los estudiantes. El modelo predictivo permitió la detección temprana e identificación de estudiantes en riesgo de desertar y así evitar el abandono de los estudiantes a través de la intervención adecuada. Además, también permitió identificar y escalar las variables de acuerdo con su peso en la deserción escolar (Chung & Lee, 2019). Otro caso fue implementado en Guatemala y Honduras, donde utilizaron datos administrativos para estimar la deserción escolar como un resultado binario para países de ingresos medios-bajos; a partir de técnicas estadísticas como probabilidad lineal; en él se combinó la información de varias variables altamente correlacionadas que miden similares características y la construcción de índices usando un Análisis de Componentes Principales; de acuerdo con las conclusiones a las que llegaron, este modelo identifica correctamente al 80 % de los estudiantes de sexto grado que abandonarán los estudios durante el próximo año escolar (Adelman et al., 2018).

Además de las técnicas de predicción empleadas en los estudiantes de las escuelas o colegios de educación primaria y secundaria ya mencionados anteriormente, también se han utilizado técnicas o modelos para los estudiantes universitarios; mediante el análisis de factores

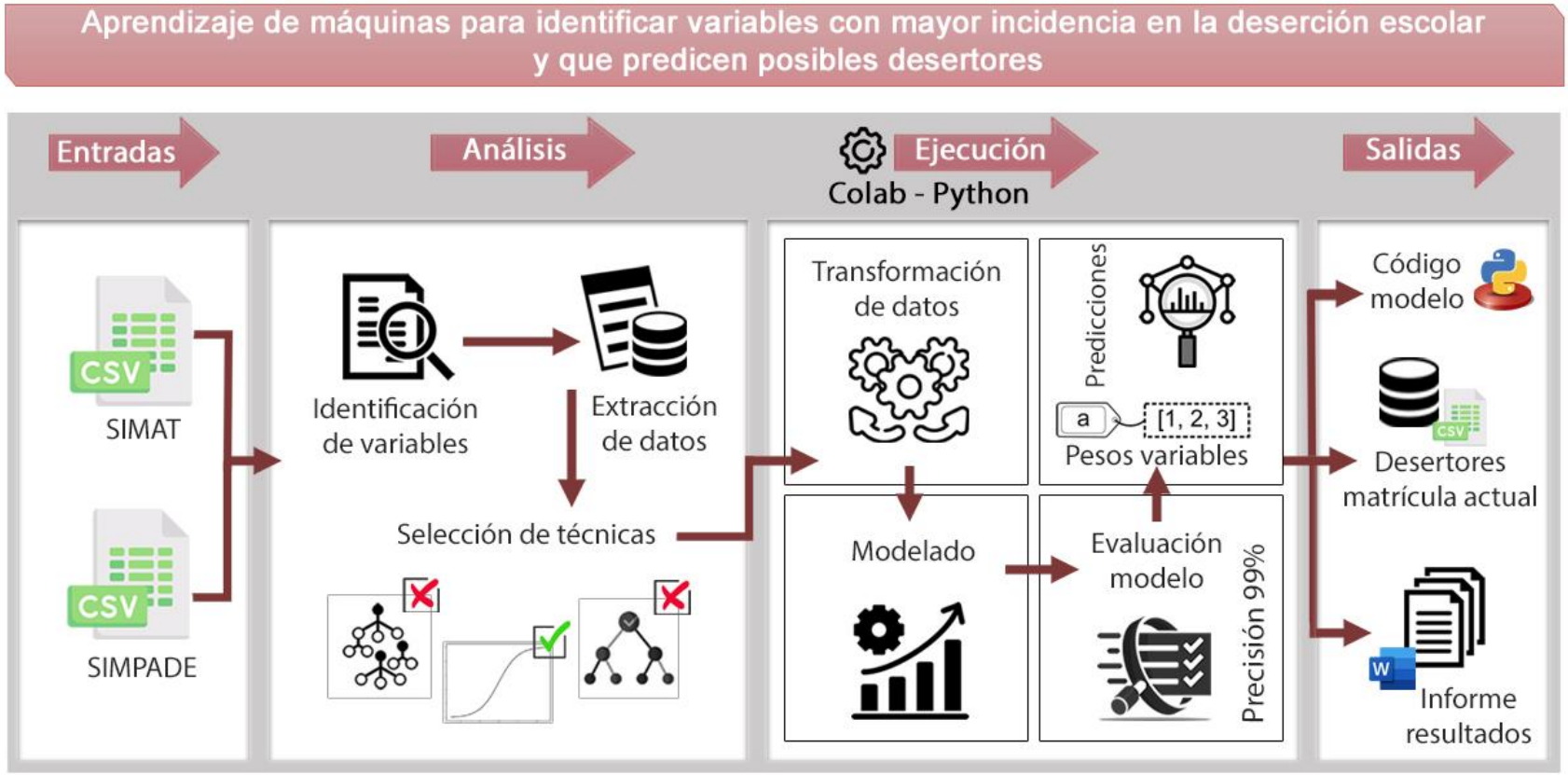
y aplicando técnicas de analítica predictiva se identifica la relación de estos y la caracterización de los estudiantes que egresan de la universidad y con esto se puede predecir el riesgo de deserción (Pasic & Kucak, 2020). Otro modelo empleado en universidades se basa en un modelo de aprendizaje automático donde se recopila datos de la asignatura de álgebra de las escuelas secundarias a los que los estudiantes han asistido antes y así se calcula la probabilidad de no terminar un programa con éxito (Aguirre & Perez, 2020).

Por otra parte, y de acuerdo con la realidad que vivimos hoy en día en cuanto a la educación en línea, la deserción de este tipo de estudiantes debe realizarse a través de una clasificación binaria, donde de acuerdo con las variables históricas del estudiante, el proveedor asigna una etiqueta de acuerdo con su futuro comportamiento. Los cinco modelos predictivos de abandono del aprendizaje en línea más comunes son la regresión logística (LR), máquinas de vectores de soporte (SVM), red neuronal (NN), árbol de decisiones (DT) y bosques aleatorios (RF) (Coussement et al., 2020).

En el 2019, Viloria et al. aplican 4 métodos para la detección de los alumnos que abandonan los estudios en educación superior; seleccionó dos métodos para generar modelos, como ellos lo llamaron, comprensibles: árboles de decisión y regresión logística; y dos métodos que ofrecen una gran capacidad de precisión: naive bayes y k-nearest neighbors. El conjunto de estos 4 métodos les arrojó, según ellos, una solución comprensible y precisa, siendo esta última evaluada principalmente por el porcentaje de abandonos detectados.

Para resumir, la aplicación de métodos estadísticos de aprendizaje para abordar la deserción escolar se centra en la utilización de métodos como Bosques Aleatorios, Redes Bayesianas, Redes Neuronales, Árboles De Decisión y Regresión Logística, Naive Bayes Y K-Neighbors Vecinos Más Cercanos, estos podrían ofrecer una solución comprensible y de precisión (Viloria et al., 2019).

8. Diagrama de Arquitectura del Modelo Construido



9. Desarrollo del Proyecto

9.1. Desarrollo del Objetivo Específico 1: Realizar la Comprensión y Preparación de los Datos Suministrados por la Secretaría de Educación de Medellín.

El desarrollo de este objetivo comprende las fases I, II y III de la metodología CRISP DM, utilizadas en el desarrollo del trabajo. Estas fases se dividen en seis actividades.

9.1.1. Fase I – Entendimiento del Negocio

Actividad 1: Examinar la situación actual de los procesos de la Secretaría de Educación en los que se desarrollan las estrategias para la prevención de deserción escolar.

Descripción de la Entidad

La Alcaldía de Medellín es la autoridad a nivel territorial encargada de administrar los recursos del Distrito Especial de Ciencia, Tecnología e Innovación de Medellín, en pro de que éstos sean utilizados para el bienestar de todos sus habitantes y del territorio. La Entidad trabaja a través de un modelo de gestión por procesos; cuenta con 10 procesos de apoyo, quienes brindan los recursos para el funcionamiento de la operación; 15 misionales, que ejecutan la operación; uno estratégico que planea la operación y uno de evaluación y control que realiza seguimiento y control a los demás procesos; para un total de 27 procesos.

Objetivos de la Entidad

La razón principal de la Alcaldía y su función dentro de la sociedad va encaminada a:

- Fomentar en conjunto con la sociedad el desarrollo humano.
- Garantizar el acceso a oportunidades y el ejercicio de los derechos fundamentales como salud y educación.

- Impulsar el crecimiento económico en el ámbito territorial.
- Promover la construcción de una ciudad segura, con espacios públicos modernos e incluyentes.
- Velar por el adecuado manejo de los recursos naturales y del ambiente.

Procesos del negocio

La Entidad cuenta con 27 procesos para el adecuado desarrollo de su misión; a continuación, se nombra cada uno de ellos:

1. Comunicación Pública
2. Gestión Cultural
3. Gestión del riesgo de desastres
4. Gestión del Desarrollo Económico
5. Gestión de la Educación
6. Evaluación y Mejora Institucional
7. Gestión Jurídica
8. Gestión Integral del Talento Humano
9. Servicio a la Ciudadanía
10. Gestión Catastral
11. Gestión de Servicios Públicos Domiciliarios y No Domiciliarios
12. Gestión del Control Urbanístico
13. Gestión de Hacienda Pública
14. Gestión Social del Riesgo
15. Gestión de la Obra Pública
16. Gestión de la Tecnología de la información y las Comunicaciones
17. Gestión Ambiental
18. Gestión de la Movilidad
19. Fortalecimiento de la Ciudadanía
20. Direccionamiento Estratégico
21. Gestión de la Información
22. Gestión de la Salud
23. Gestión de la Gobernanza Local
24. Gestión de la Seguridad
25. Administración de Bienes Muebles e Inmuebles
26. Gestión de Compras Públicas Transparentes
27. Mantenimiento de Bienes Muebles e Inmuebles

Figura 3. *Modelo de operación por procesos Alcaldía de Medellín*

Nota: Tomado de *Listado Maestro de Documentos*, 2022, ISOlución Alcaldía de Medellín

Proceso específico donde está la problemática

A partir del uso de bases de datos de matrícula de las instituciones educativas del Distrito de Medellín, se pretende realizar un modelo predictivo por medio de técnicas supervisadas con el objetivo de brindar información al proceso de Gestión de la Educación que le permita prevenir y disminuir la deserción escolar y fortalecer las estrategias de permanencia estudiantil que tiene la Entidad. Gestión de la Educación es un proceso misional liderado por la Secretaría de Educación de Medellín; el proceso cuenta con 44 actividades que en conjunto tienen por objetivo garantizar la prestación del servicio educativo en las comunas y corregimientos del Distrito Especial de Ciencia, Tecnología e Innovación de Medellín; a través de las políticas y estrategias de acceso, permanencia, calidad y pertinencia, la ejecución de asesoría y asistencia técnica, inspección, vigilancia y control a los establecimientos educativos del distrito; para promover la articulación de la educación básica y media con la educación superior.

El alcance del proceso inicia con la planeación de la oferta, la demanda, el acceso y la permanencia en el sistema educativo, hasta la educación media formal regular y educación para el trabajo y el desarrollo humano; continúa con la gestión de la calidad del servicio educativo, la Inspección, vigilancia y control, asesoría y asistencia técnica a establecimientos educativos, la promoción de la articulación de éstas con la educación superior, y termina con las acciones de mejora del proceso.

Una de las actividades del proceso de Gestión de la Educación, en cabeza de la Subsecretaría de Planeación Educativa; es “asegurar el acceso y la permanencia de las niñas, niños, jóvenes, adolescentes y adultos en el servicio educativo; mediante la ejecución de estrategias de acceso y permanencia al Servicio Educativo”. Esta actividad está descrita en el procedimiento “PR-EDUC-068 -Implementación de estrategias de acceso y permanencia”, su

objetivo es identificar, implementar y realizar seguimiento a las estrategias de acceso y permanencia que permitan el ingreso y continuidad en el servicio educativo de la ciudad.

Tabla 2. *Descripción PR-EDUC-068 –Implementación de estrategias de acceso y permanencia*

N°	TAREA
1	Nombre Tarea: revisar los lineamientos ministeriales y del ente territorial sobre acceso y permanencia:
	(Responderá al qué, cómo, cuándo, con qué): En el último trimestre del año los profesionales de acceso, cobertura y permanencia consultan la siguiente información: • En la página del Ministerio de Educación Nacional los lineamientos sobre acceso y permanencia escolar y las comunicaciones remitidas por esta entidad a la SEM. • Los informes que se elaboren desde análisis sectorial sobre acceso y permanencia y los reportes de los sistemas SIMPADE y SIMAT, revisados y validados en comité de permanencia. • Los lineamientos relacionados con las estrategias de acceso y permanencia que emita la Secretaría de Educación para la atención a la primera infancia. Una vez sean revisados los documentos se identifican las estrategias y/o acciones de acceso y permanencia que la Secretaría de Educación de Medellín implementará.
	Responsable: PROFESIONAL UNIVERSITARIO, CONTRATISTA
	Área: SECRETARÍA DE EDUCACIÓN
	Oficina: SUBS. DE PLANEACIÓN EDUCATIVA
	Registro: memorias y actas comité de permanencia
2	Nombre Tarea: identificar las necesidades de acceso y permanencia
	(Responderá al qué, cómo, cuándo, con qué): los profesionales de acceso, cobertura y permanencia juntamente con las Instituciones Educativas (IE) y el grupo de educación Inicial identifica las necesidades a partir de: • Resultado de las caracterizaciones e informes generados en SIMPADE • Solicitudes de la familia para el acceso al servicio educativo o a beneficios complementarios en las Instituciones Educativas. • Análisis de información del ente territorial (históricos de ejecución de las estrategias, análisis de las acciones emprendidas por el ente territorial en comité de permanencia). • Solicitudes directas de las IE direcciones de núcleo, entidades prestadoras del servicio.

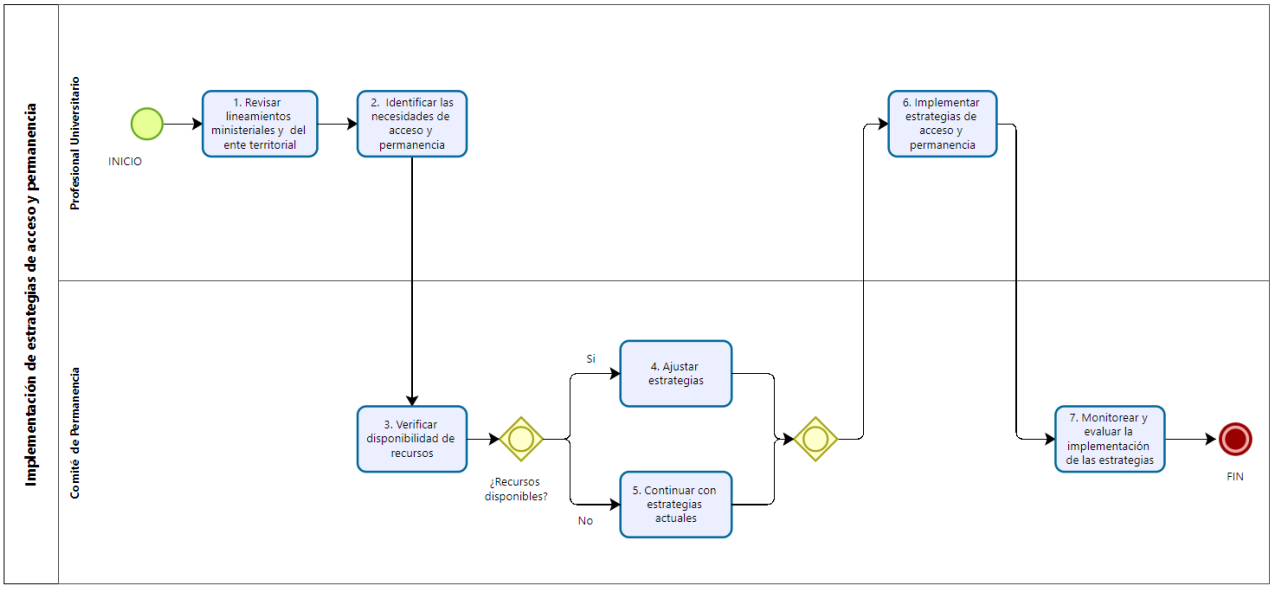
N°	TAREA
	• Articulación de la Secretaría con otras dependencias de la administración distrital.
	Responsable: PROFESIONAL UNIVERSITARIO, CONTRATISTA
	Área: SECRETARÍA DE EDUCACIÓN
	Oficina: Subsecretaría de Planeación Educativa, Subsecretaría de Prestación del Servicio Educativo
	Registro: memorias comité de permanencia
3	Nombre Tarea: verificar la disponibilidad de recursos.
	(Responderá al qué, cómo, cuándo, con qué): el comité de permanencia con el fin de ajustar las estrategias de acceso y permanencia verifica la disponibilidad de recursos físicos y financieros teniendo en cuenta lo siguiente: • Las características del territorio • Capacidad instalada de las instituciones educativas y entidades prestadoras del servicio educativo, infraestructura y talento humano • Disponibilidad presupuestal • Pertinencia de la estrategia para dar respuesta a las necesidades
	Responsable: COMITÉ DE PERMANENCIA
	Área: SECRETARIA DE EDUCACION
	Oficina: Subsecretaría de Planeación Educativa, Subsecretaría de Prestación del Servicio Educativo
	Registro: memorias comité de permanencia
4	Nombre Tarea: Ajustar y planear la implementación de las estrategias de acceso y permanencia acorde a las necesidades identificadas.
	(Responderá al qué, cómo, cuándo, con qué): con base en las necesidades identificadas, la viabilidad de recursos y la información revisada sobre las estrategias de acceso y permanencia, los profesionales de acceso, cobertura y permanencia de manera conjunta con los responsables de los programas y proyectos en ejecución desde la Prestación del Servicio Educativo, analizan los factores de riesgos de deserción identificados por las instituciones educativas para priorizarlas y ajustar las estrategias e iniciar la implementación.
	Responsable: COMITÉ DE PERMANENCIA
	Área: SECRETARIA DE EDUCACION

N°	TAREA
	Oficina: Subsecretaría de Planeación Educativa, Subsecretaría de Prestación del Servicio Educativo
	Registro: memorias comité de permanencia
5	Nombre Tarea: continuar con las estrategias actuales.
	(Responderá al qué, cómo, cuándo, con qué): después de la verificación de recursos el comité informa sobre la continuidad de las estrategias actuales sin ajustes autorizados.
	Responsable: COMITÉ DE PERMANENCIA
	Área: SECRETARIA DE EDUCACION
	Oficina: Subsecretaría de Planeación Educativa, Subsecretaría de Prestación del Servicio Educativo
	Registro: Memorias comité de permanencia
6	Nombre Tarea: implementar las estrategias de acceso y permanencia
	(Responderá al qué, cómo, cuándo, con qué): los profesionales de acceso, cobertura y permanencia ponen en ejecución las estrategias de acceso y permanencia conforme a lo establecido en la planeación descrita en la actividad 3 de este documento, y teniendo en cuenta los procedimientos y/o documentos guía definidos para cada una de ellas: Acceso: • Ruta de escolarización • Búsqueda activa de desescolarizados Permanencia: • Programa de Alimentación Escolar – PAE • Transporte escolar • Modelos Educativos Flexibles • Control de asistencia, retiros y deserción – Rutas para la implementación. Adicionalmente, se articulan programas y proyectos que en clave de acceso y permanencia son implementadas desde la subsecretaría de prestación del servicio educativo.
	Responsable: PROFESIONAL UNIVERSITARIO, CONTRATISTA
	Área: SECRETARIA DE EDUCACION

N°	TAREA
	Oficina: SUBS. DE PLANEACIÓN EDUCATIVA
	Registro: actas de ejecución
7	Nombre Tarea: monitorear y evaluar la implementación de las estrategias de acceso y permanencia
	(Responderá al qué, cómo, cuándo, con qué): en el comité de permanencia periódicamente se evalúan los resultados de las estrategias de acceso y permanencia implementadas por la Secretaría. La evaluación verifica la efectividad de la estrategia con referencia al acceso, incremento de la atención, la permanencia escolar y la reducción de factores de riesgo que afectan la prestación del servicio educativo y se toman acciones a que haya lugar.
	Responsable: COMITÉ DE PERMANENCIA
	Área: SECRETARÍA DE EDUCACIÓN
	Oficina: Subsecretaría de Planeación Educativa, Subsecretaría de Prestación del Servicio Educativo
	Registro: acta comité de permanencia y plan de mejoramiento.

Nota: Tomado de *Listado Maestro de Documentos*, 2019, ISOlución Alcaldía de Medellín

Figura 4. *Flujograma del procedimiento PR-EDUC-068*



Aplicación y análisis de Machine Learning

La aplicación de técnicas de aprendizaje automático dentro de este estudio tiene los siguientes propósitos e indicadores para contribución a su desarrollo y análisis:

Propósitos u objetivos de analítica

- Identificar las variables que tienen mayor correlación con la deserción estudiantil.
- Generar un modelo que permita predecir los posibles desertores de las instituciones educativas.

Indicadores o métricas para análisis

- Total de posibles desertores generados por el modelo/ Total de registros del modelo
- Total posibles desertores por grado escolar

9.1.2. Fase II. Estudio y Comprensión de los Datos

Actividad 2: Conseguir las bases de datos con las variables necesarias determinadas.

Bases de datos iniciales

Las bases de datos necesarias para la investigación fueron obtenidas a través de la Secretaría de Educación de Medellín, generadas por medio de los sistemas de información nacionales (externos), Sistema Integrado de Matrícula (SIMAT) y Sistema de Información para el Monitoreo, la Prevención y el Análisis de la Deserción Escolar (SIMPADE); cabe anotar que para hacer uso de estas bases de datos que contienen información sensible de estudiantes menores de edad, se tramitó ante la Secretaría de Educación permiso de uso y confidencialidad de datos (ver anexo A) donde se fijan los términos y condiciones bajo los cuales se mantendrá la información que será suministrada de forma anonimizada.

Las bases generadas y suministradas están en formato .csv, en ellas se encuentra información del estudiante que permite caracterizarlo en temas relacionados con la permanencia escolar y el índice de riesgo de deserción, tales como embarazos, discapacidades, abandonos escolares, repitencias, deserción, víctimas de discriminación, agresiones y conflictos, entre otros.

Para evaluar el núcleo familiar, se tienen datos como tipo y tenencia de vivienda, nivel de escolaridad del acudiente, número de personas que conforman el hogar y acompañamiento por parte del acudiente al desempeño escolar.

En cuanto a las variables relacionadas con la institución educativa, hay información básica como nombre de la institución, sede, código DANE y complementaria como carácter de la institución, tipo de calendario, zona de atención, entre otras, que están relacionadas con contexto tanto interno como externo.

Finalmente se tienen datos referentes a la información básica del municipio y a acciones adelantadas para garantizar la permanencia escolar y el acceso a estas por parte de los estudiantes como nombres y números de estrategias a las que ha accedido.

Figura 5. *Listado de variables del Sistema Integrado de Matrícula (SIMAT)*

```
[14] simat = pd.read_csv('/content/drive/MyDrive/Trabajo_grado/Datos/SIMAT_MEDELLIN_2021.csv')
```

```
simat.info ()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 425999 entries, 0 to 425998
Data columns (total 59 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   ANNO_INF                    425999 non-null  int64
 1   MUN_CODIGO                  425999 non-null  int64
 2   CODIGO_DANE                 425999 non-null  float64
 3   CODIGO_DANE_SEDE            425999 non-null  float64
 4   CONS_SEDE                   425999 non-null  float64
 5   ESTRATO                     425999 non-null  int64
 6   SISBEN                      245804 non-null  float64
 7   FECHA_NACIMIENTO            425999 non-null  object
 8   GENERO                      425999 non-null  object
 9   POB_VICT_CONF               425999 non-null  int64
 10  DPTO_EXP                    15102 non-null   float64
 11  MUN_EXP                     15102 non-null   float64
 12  PROVIENE_SECTOR_PRIV        425999 non-null  object
 13  PROVIENE_OTRO_MUN           425999 non-null  object
 14  TIPO_DISCAPACIDAD           425999 non-null  int64
 15  CAP_EXC                     425999 non-null  int64
 16  ETNIA                       425999 non-null  int64
 17  RES                         425999 non-null  int64
 18  TIPO_JORNADA                425999 non-null  int64
 19  CARACTER                    425999 non-null  int64
 20  ESPECIALIDAD                425999 non-null  int64
 21  GRADO                       425999 non-null  int64
 22  GRUPO                       410985 non-null  float64
 23  METODOLOGIA                 425999 non-null  int64
 24  SUBSIDIADO                  425999 non-null  object
 25  REPITENTE                   425999 non-null  object
 26  NUEVO                       425999 non-null  object
 27  SIT_ACAD_ANO_ANT            425999 non-null  int64
 28  CON_ALUM_ANO_ANT            425999 non-null  int64
 29  FUE_RECU                    425999 non-null  int64
 30  ZON_ALU                     425999 non-null  int64
 31  CAB_FAMILIA                 425999 non-null  object
 32  BEN_MAD_FLIA                425999 non-null  object
 33  BEN_VET_FP                  425999 non-null  object
 34  BEN_HER_NAC                 425999 non-null  object
 35  INTERNADO                   337428 non-null  float64
 36  VAL_DES_PERIODO1            425999 non-null  object
 37  VAL_DES_PERIODO2            425999 non-null  object
 38  NUM_CONVENIO                35837 non-null   float64
 39  SEDE_ID                     425999 non-null  int64
 40  EST_ID                      425999 non-null  int64
 41  CODIGO_SED                  425999 non-null  int64
 42  DPTO_CARGA                  425999 non-null  int64
 43  NOMBRE_ESTABLECIMIENTO      425999 non-null  object
 44  CTE_ID_SECTOR               425999 non-null  int64
 45  CTE_ID_CALENDARIO           425999 non-null  int64
 46  NOMBRE_SEDE                 425999 non-null  object
 47  CTE_ID_ZONA                 425999 non-null  int64
 48  ESTADO_DEFINITIVO           425999 non-null  int64
 49  Divipola_MUNICIPIO          425999 non-null  int64
 50  EDAD                        425999 non-null  int64
 51  SECTOR_CONPES               425999 non-null  int64
 52  DISCAPACIDAD1               425999 non-null  int64
```

Figura 6. *Listado de variables del Sistema de Información para el Monitoreo, la Prevención y el Análisis de la Deserción Escolar (SIMPADE)*

```
simpade = pd.read_csv('/content/drive/MyDrive/Trabajo_grado/Datos/Simpade112020.csv')
```

```
simpade.info (5)
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 263075 entries, 0 to 263074
Data columns (total 158 columns):
 #    Column                                                Dtype
---   ------                                                -----
 0    SPD_COD_SED                                           int64
 1    SPD_SECRETARIA                                        object
 2    SPD_COD_DEPTO                                         int64
 3    SPD_DEPTO                                             object
 4    SPD_COD_MUN                                           int64
 5    SPD_NOMBRE_NUM                                        object
 6    SPD_DANE_EE                                           float64
 7    SPD_NOMBRE_EE                                         object
 8    SPD_DANE_SEDE                                         float64
 9    SPD_NOMBRE_SEDE                                       object
 10   VIVE_SOLO                                             object
 11   EMBARAZO                                              object
 12   TRABAJA                                               object
 13   MAT_PAT_TEMPRANA_EDAD                                 object
 14   VICT_DISCRI_LGBTI                                     object
 15   VICTIMA_AGRESIONES                                    object
 16   VICTIMA_DISCRIMINACION_CODIGO                         object
 17   VICTIMA_DISCRIMINACION_NOMBRE                         object
 18   VICTIMA_DISCRIMINACION_OTRAS                          object
 19   CONTEO_DISCRIMINACION                                 int64
 20   QUE_DESEA_ESTUDIAR                                    object
 21   NUMERO_PERSONAS_HOGAR                                 float64
 22   TIPO_VIVIENDA                                         object
 23   TENENCIA_VIVIENDA                                     object
 24   SERVIC_PUBLICO_CODIGO                                 object
 25   SERVIC_PUBBLICO_NOMBRE                                object
 26   CUENTA_SERVICIOS                                      int64
 27   EDUCACION_PREESCOLAR                                  object
 28   ANOS_EDUCACION_PREESCOLAR                             float64
 29   RET_EST_SIN_TERMINAR_ANO_ESC                          object
 30   ABANDONOS_TEMPORALES                                  float64
 31   REPETIDA_ANO                                          object
 32   VECES_REPETIDO_ANO                                    float64
 33   REPITIENDO_GRADO_ACTUAL                               object
 34   ANTECED_DICIPL_VIDA_ACADEM                            object
 35   ASIST_PROM_ANO_ANTERIOR                               object
 36   DIF_APREN_DIAG_ESPEC                                  object
 37   VINCUL_MODAL_EDUC_INICIAL                             object
 38   ASIGNA_NO_ESTA_APROB_COD_P1                           object
 39   ASIGNA_NO_ESTA_APROB_NOMBRE_P1                        object
 40   ASIGNA_NO_ESTA_APROB_CONT_P1                          int64
 41   ASIGNA_NO_ESTA_APROB_COD_P2                           object
 42   ASIGNA_NO_ESTA_APROB_NOMBRE_P2                        object
 43   ASIGNA_NO_ESTA_APROB_CONT_P2                          int64
 44   NOM_ESTRAT_COD                                        object
 45   NOM_ESTRAT_NOMBRE                                     object
 46   NOM_ESTRAT_CONTEO                                     int64
 47   PARENTEZCO_CODIGO                                     object
 48   PARENTEZCO_NOMBRE                                     object
 49   NIVEL_EDUCATIVO_MEDIA_COMPLETA                        float64
 50   NIVEL_EDUCATIVO_MEDIA_INCOMPLETA                      float64
 51   NIVEL_EDUCATIVO_POSGRADO                              float64
 52   NIVEL_EDUCATIVO_PRIMARIA_COMPLETA                     float64
 53   NIVEL_EDUCATIVO_PRIMARIA_INCOMPLETA                   float64
 54   NIVEL_EDUCATIVO_SIN_EDUCACION                         float64
 55   NIVEL_EDUCATIVO_SUPERIOR_COMPLETA                     float64
 56   NIVEL_EDUCATIVO_SUPERIOR_INCOMPLETA                   float64
 57   ASISTE_REU_DIF_BOLETIN_ALGUNAS_VECES                  float64
 58   ASISTE_REU_DIF_BOLETIN_CASI_NUNCA                     float64
 59   ASISTE_REU_DIF_BOLETIN_CASI_SIEMPRE                   float64
 60   ASISTE_REU_DIF_BOLETIN_NUNCA                          float64
 61   ASISTE_REU_DIF_BOLETIN_SIEMPRE                        float64
 62   ASISTE_ENTREGA_INFORMES_ALGUNAS_VECES                 float64
 63   ASISTE_ENTREGA_INFORMES_CASI_NUNCA                    float64
 64   ASISTE_ENTREGA_INFORMES_CASI_SIEMPRE                  float64
 65   ASISTE_ENTREGA_INFORMES_NUNCA                         float64
 66   ASISTE_ENTREGA_INFORMES_SIEMPRE                       float64
 67   TIPO_EMPLEO_NO_TIENE                                  float64
 68   TIPO_EMPLEO_PERMANENTE                                float64
 69   TIPO_EMPLEO_TEMPORAL                                  float64
 70   FRECUENCIA_CAMBIO_DOMIC_ENTRE_1_Y_2_VECES             float64
 71   FRECUENCIA_CAMBIO_DOMIC_ENTRE_3_Y_4_VECES             float64
 72   FRECUENCIA_CAMBIO_DOMIC_MAS_DE_4_VECES                float64
 73   FRECUENCIA_CAMBIO_DOMIC_NO_HA_CAMBIADO                float64
 74   ANNO_INF                                              int64
 75   APOYO ACADEMICO                                       object
 76   SRPA                                                  object
 77   CODIGO_SED                                            int64
 78   SECRETARIA                                            object
 79   DEPTO                                                 object
 80   MUN_CODIGO                                            int64
 81   Divipola_MUNICIPIO                                    object
 82   DANE_EE                                               float64
 83   NOMBRE_ESTABLECIMIENTO                                object
 84   DANE_SEDE                                             float64
 85   NOMBRE_SEDE                                           object
 86   CONS_SEDE                                             float64
 87   PER_ID                                                int64
 88   DEPTO_DE_EXPEDICION_DOC                               object
 89   MUN_DE_EXPEDICION_DOC                                 object
 90   RES_DEPTO                                             object
 91   RES_MUN                                               object
 92   ESTRATO                                               object
 93   SISBEN                                                object
 94   FECHA_NACIMIENTO                                      object
 95   NAC_DEPTO                                             object
 96   NAC_MUN                                               object
 97   GENERO                                                object
 98   POBLACION_VICTIMA_CONFLICTO                           object
 99   DPTO_EXP                                              object
 100  MUN_EXP                                               object
 101  PROVIENE_DEL_SECTOR_PRIVADO                           object
 102  PROVIENE_DE_OTRO_MUNICIPIO                            object
 103  TIPO_DE_DISCAPACIDAD                                  float64
 104  CAPACIDADES_EXCEPCIONALES                             float64
 105  ETNIA                                                 object
 106  RESGUARDO                                             object
 107  INS_FAMILIAR                                          object
 108  JORNADA                                               object
 109  car?cter                                              object
 110  ESPECIALIDAD                                          object
 111  GRADO                                                 object
 112  GRUPO                                                 object
 113  METODOLOGIA                                           object
 114  SUBSIDIADO                                            object
 115  REPITENTE                                             object
 116  NUEVO                                                 object
 117  SITUACION_ACADEMICA_AÑO_ANTERIOR                      object
 118  CONDICION_DEL_ALUMNO_AL_FINALIZAR_EL_AÑO_ANTERIOR     object
 119  FUE_RECU                                              object
 120  ZONA__ALU                                             object
 121  MADRE_CABEZA_DE_FAMILIA                               object
 122  BENEF._MADRE_CABEZA_DE_FLIA                           object
 123  BENEF._VETERANOS                                      object
 124  BENEF.HERORES                                         object
 125  INTERNADO                                             object
 126  NUM_CONVENIO                                          float64
 127  SEDE_ID                                               int64
 128  ESTADOSEDE                                            float64
 129  EST_ID                                                int64
 130  SECTOR                                                object
 131  CALENDARIO                                            object
 132  ZONA_DE_ATENCIÓN                                      object
 133  EDAD                                                  int64
 134  NIVEL_CONPES                                          object
 135  MES_FINAL                                             int64
 136  MESES_ATENCION_6A                                     int64
 137  ESTADO_CONSOLIDADO                                    int64
 138  04_ATENDIDO                                           float64
 139  04_CODIGO_SED                                         float64
 140  05_ATENDIDO                                           float64
 141  05_CODIGO_SED                                         float64
 142  06_ATENDIDO                                           float64
 143  06_CODIGO_SED                                         float64
 144  07_ATENDIDO                                           float64
 145  07_CODIGO_SED                                         float64
 146  08_ATENDIDO                                           float64
 147  08_CODIGO_SED                                         float64
 148  09_ATENDIDO                                           float64
 149  09_CODIGO_SED                                         float64
 150  10_ATENDIDO                                           float64
 151  10_CODIGO_SED                                         float64
 152  11_ATENDIDO                                           int64
 153  11_CODIGO_SED                                         int64
 154  COD_SED                                               int64
 155  CONPES_POB_ATENDIDA                                   float64
 156  INDICE_DE_DESERCION                                   float64
 157  FECHA_CORTE                                           object
dtypes: float64(54), int64(21), object(83)
memory usage: 317.1+ MB
```

Actividad 3: Identificar los problemas que las bases tengan en cuanto a la calidad de los datos.

Calidad de los datos

Completitud: toda la información para realizar la investigación se encuentra disponible en la Secretaría de Educación de Medellín, a través de los sistemas SIMAT y SIMPADE; en cuanto a los datos, hay algunos faltantes, pero en su mayoría estas variables no son relevantes en el desarrollo de la investigación.

Exactitud: la información suministrada tiene un alto porcentaje de ser correcta; pues es diligenciada directamente por las instituciones con la información de sus estudiantes; sin embargo, no está 100% libre de error, porque pueden presentarse errores de digitación por ser la información ingresada manualmente.

Conformidad: los valores de los siguientes atributos no están conforme a su formato numérico: Dane_EE, Spd_Dane_EE, Spd_Dane_Sede, Dane_Sede, Cons_Sede; todos los demás datos están conformes con los formatos esperados.

Oportunidad: la información se obtiene anualmente, dado que las bases de datos son generadas por el Ministerio de Educación Nacional a partir de la información recopilada por las instituciones educativas durante el periodo de un año académico.

Duplicidad: la información suministrada contiene algunas variables que hacen referencia a los mismos objetos de datos, ellas son; Victima_Discriminacion_Codigo, Victima_Discriminacion_Nombre, Codigo_Sed, Secretaría, Depto, Mun_Codigo, Divipola_Municipio, Dane_EE, Nombre_Establecimiento, Dane_Sede, Nombre_Sede, Cons_Sede, Fecha_Nacimiento, Edad.

Consistencia: la información contenida en las bases no contiene datos que proporcionen información diferente sobre el mismo objeto de datos.

Integridad: la conectividad de los datos es clara al igual que las relaciones con otros datos; sin embargo, las siguientes variables tienen una relación importante con el estudio, pero contienen gran cantidad de datos faltantes, algunos de ellos pueden ser suplidos al integrar las bases de datos; Numero_Personas_Hogar, Abandonos_Temporales, Veces_Repetido_Ano, Repitiendo_Grado_Actual, Anteced_Dicipl_Vida_Academ, Asist_Prom_Ano_Anterior, Tipo_De_Discapacidad y Capacidades_Excepcionales.

De acuerdo con el análisis de calidad de los datos se eliminan las columnas que no se consideran relevantes dentro del estudio.

Figura 7. *Lista de variables eliminadas*

```
#Eliminar Columnas Irrelevantes
simpade.drop(['SPD_COD_SED', 'SPD_SECRETARIA', 'SPD_COD_DEPTO',  'SPD_DEPTO', 'SPD_COD_MUN', 'SPD_NOMBRE_NUM', 'SPD_DANE_EE', 'SPD_NOMBRE_EE', 'SPD_DANE_SEDE', 'SPD_NOMBRE_SEDE',
'VICTIMA_DISCRIMINACION_CODIGO', 'VICTIMA_DISCRIMINACION_NOMBRE', 'VICTIMA_DISCRIMINACION_OTRAS', 'CONTEO_DISCRIMINACION', 'QUE_DESEA_ESTUDIAR', 'TIPO_VIVIENDA',
'SERVIC_PUBLICO_CODIGO', 'SERVIC_PUBBLICO_NOMBRE', 'CUENTA_SERVICIOS', 'EDUCACION_PREESCOLAR', 'ANOS_EDUCACION_PREESCOLAR', 'VINCUL_MODAL_EDUC_INICIAL', 'ASIGNA_NO_ESTA_APROB_COD_P1
'ASIGNA_NO_ESTA_APROB_NOMBRE_P1', 'ASIGNA_NO_ESTA_APROB_CONT_P1', 'ASIGNA_NO_ESTA_APROB_COD_P2', 'ASIGNA_NO_ESTA_APROB_NOMBRE_P2', 'ASIGNA_NO_ESTA_APROB_CONT_P2',
'NOM_ESTRAT_COD', 'PARENTEZCO_CODIGO', 'PARENTEZCO_NOMBRE', 'NIVEL_EDUCATIVO_MEDIA_COMPLETA', 'NIVEL_EDUCATIVO_MEDIA_INCOMPLETA', 'NIVEL_EDUCATIVO_POSGRADO',
'NIVEL_EDUCATIVO_PRIMARIA_COMPLETA', 'NIVEL_EDUCATIVO_PRIMARIA_INCOMPLETA', 'NIVEL_EDUCATIVO_SIN_EDUCACION', 'NIVEL_EDUCATIVO_SUPERIOR_COMPLETA',
'NIVEL_EDUCATIVO_SUPERIOR_INCOMPLETA', 'ANNO_INF', 'SRPA', 'CODIGO_SED', 'SECRETARIA', 'DEPTO', 'MUN_CODIGO', 'Divipola_MUNICIPIO', 'DANE_EE', 'NOMBRE_ESTABLECIMIENTO',
'DANE_SEDE', 'NOMBRE_SEDE', 'CONS_SEDE', 'DEPTO_DE_EXPEDICION_DOC', 'MUN_DE_EXPEDICION_DOC', 'RES_DEPTO', 'RES_MUN', 'FECHA_NACIMIENTO', 'NAC_DEPTO', 'NAC_MUN',
'DPTO_EXP', 'MUN_EXP', 'PROVIENE_DEL_SECTOR_PRIVADO', 'PROVIENE_DE_OTRO_MUNICIPIO', 'RESGUARDO', 'INS_FAMILIAR', 'JORNADA', 'car?cter', 'ESPECIALIDAD', 'GRUPO',
'SUBSIDIADO', 'NUEVO', 'FUE_RECU', 'MADRE_CABEZA_DE_FAMILIA', 'BENEF._VETERANOS', 'BENEF.HEROES', 'INTERNADO', 'NUM_CONVENIO', 'SEDE_ID', 'ESTADOSEDE', 'EST_ID',
'SECTOR', 'CALENDARIO', 'NIVEL_CONPES', 'MES_FINAL', 'MESES_ATENCION_6A', 'ESTADO_CONSOLIDADO', '04_ATENDIDO', '04_CODIGO_SED', '05_ATENDIDO', '05_CODIGO_SED',
'06_ATENDIDO', '06_CODIGO_SED', '07_ATENDIDO', '07_CODIGO_SED', '08_ATENDIDO', '08_CODIGO_SED', '09_ATENDIDO', '09_CODIGO_SED', '10_ATENDIDO', '10_CODIGO_SED',
'11_ATENDIDO', '11_CODIGO_SED', 'COD_SED', 'CONPES_POB_ATENDIDA', 'INDICE_DE_DESERCION', 'FECHA_CORTE'], axis='columns', inplace=True)
```

Actividad 4: Generar informe con la descripción de los datos.

Descripción de los datos

Después de seleccionar las variables relevantes para el estudio de acuerdo con las bases de datos obtenidas, el diccionario de datos resultante es el siguiente**:**

Tabla 3. *Diccionario de datos*

ATRIBUTO	DESCRIPCIÓN	TIPO	NATURALEZA	ESCALA	OBSERVACION (di-poli) o (continua-discontinua
ABANDONOS_TEMPORALES	Presenta abandonos temporales	Categórica	Cualitativa	Nominal	Politómico
ANTECED_DICIPL_VIDA_ACADEM	Antecedentes disciplinarios en la vida académica	Categórica	Cualitativa	Nominal	Politómico
APOYO ACADEMICO	Apoyo académico especial para el estudiante	Categórica	Cualitativa	Nominal	Politómico
ASIST_PROM_ANO_ANTERIOR	Asistencia promedio año anterior	Categórica	Cualitativa	Nominal	Politómico
ASISTE_ENTREGA_INFORMES_ALGUNAS_VECES	Asistencia a entrega de informes - algunas veces	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_ENTREGA_INFORMES_CASI_NUNCA	Asistencia a entrega de informes - casi nunca	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_ENTREGA_INFORMES_CASI_SIEMPRE	Asistencia a entrega de informes - casi siempre	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_ENTREGA_INFORMES_NUNCA	Asistencia a entrega de informes - nunca	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_ENTREGA_INFORMES_SIEMPRE	Asistencia a entrega de informes - siempre	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_REU_DIF_BOLETIN_ALGUNAS_VECES	Asistencia a reuniones de boletín - algunas veces	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_REU_DIF_BOLETIN_CASI_NUNCA	Asistencia a reuniones de boletín - casi nunca	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_REU_DIF_BOLETIN_CASI_SIEMPRE	Asistencia a reuniones de boletín - casi siempre	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_REU_DIF_BOLETIN_NUNCA	Asistencia a reuniones de boletín - nunca	Categórica	Cualitativa	Ordinal	Politómico
ASISTE_REU_DIF_BOLETIN_SIEMPRE	Asistencia a reuniones de boletín - siempre	Categórica	Cualitativa	Ordinal	Politómico
BENEF._MADRE_CABEZA_DE_FLIA	Estudiante con madre cabeza de familia	Categórica	Cualitativa	Nominal	Dicotómico
CAPACIDADES_EXCEPCIONALES	Posee capacidades excepcionales	Categórica	Cualitativa	Nominal	Dicotómico
CONDICION_DEL_ALUMNO_AL_FINALIZAR_EL_AÑO_ANTERIOR	Condición del alumno al terminar el año anterior	Categórica	Cualitativa	Nominal	Dicotómico
DIF_APREN_DIAG_ESPEC	Aprendizaje diferenciado por tener diagnótico especial	Categórica	Cualitativa	Nominal	Dicotómico
EDAD	Edad del estudiante	Numérica	Cuantitativa	Razón	Discreta
EMBARAZO	Muestra si ha tenido embarazos	Categórica	Cualitativa	Nominal	Dicotómico
ESTRATO	Estrato socioeconómico estudiante	Categórica	Cualitativa	Ordinal	Politómico
ETNIA	Tipo de etnia	Categórica	Cualitativa	Nominal	Dicotómico
FRECUENCIA_CAMBIO_DOMIC_ENTRE_1_Y_2_VECES	Frecuencia de cambio de domicilio - 1 a 2 veces	Categórica	Cualitativa	Ordinal	Politómico
FRECUENCIA_CAMBIO_DOMIC_ENTRE_3_Y_4_VECES	Frecuencia de cambio de domicilio - 3 a 4 veces	Categórica	Cualitativa	Ordinal	Politómico
FRECUENCIA_CAMBIO_DOMIC_MAS_DE_4_VECES	Frecuencia de cambio de domicilio - más de 4 veces	Categórica	Cualitativa	Ordinal	Politómico
FRECUENCIA_CAMBIO_DOMIC_NO_HA_CAMBIADO	Frecuencia de cambio de domicilio - no ha cambiado	Categórica	Cualitativa	Ordinal	Politómico
GENERO	Género	Categórica	Cualitativa	Nominal	Dicotómico
GRADO	Grado que cursa	Categórica	Cualitativa	Ordinal	Politómico
MAT_PAT_TEMPRANA_EDAD	Presenta maternidad o paternidad a temprana edad	Categórica	Cualitativa	Nominal	Dicotómico
METODOLOGIA	Metodología de estudio	Categórica	Cualitativa	Nominal	Politómico
NOM_ESTRAT_CONTEO	Total estrategias a las que accede	Numérica	Cuantitativa	Razón	Discreta
NOM_ESTRAT_NOMBRE	Nombre de estrategia de permanencia	Categórica	Cualitativa	Nominal	Politómico

ATRIBUTO	DESCRIPCIÓN	TIPO	NATURALEZA	ESCALA	OBSERVACION (di-poli) o (continua-discontinua
NUMERO_PERSONAS_HOGAR	Número de personas que conforman el hogar	Numérica	Cuantitativa	Razón	Discreta
PER_ID	Identificador único de cada registro de estudiante en el sistema	Numérica	Cuantitativa	Razón	Discreta
POBLACION_VICTIMA_CONFLICTO	Población víctima de conflicto	Categórica	Cualitativa	Nominal	Dicotómico
REPETIDA_ANO	Repitencia de años	Categórica	Cualitativa	Nominal	Dicotómico
REPITENTE	Muestra si es un estudiante repitente	Categórica	Cualitativa	Nominal	Dicotómico
REPITIENDO_GRADO_ACTUAL	Está repitiendo grado actual	Categórica	Cualitativa	Nominal	Dicotómico
RET_EST_SIN_TERMINAR_ANO_ESC	Retiro estudiante sin terminar año escolar	Categórica	Cualitativa	Nominal	Dicotómico
SISBEN	Nivel del Sisbén	Categórica	Cualitativa	Ordinal	Politómico
SITUACION_ACADEMICA_AÑO_ANTERIOR	Situación académica del año anterior	Categórica	Cualitativa	Nominal	Dicotómico
TENENCIA_VIVIENDA	Tenencia de vivienda	Categórica	Cualitativa	Nominal	Politómico
TIPO_DE_DISCAPACIDAD	Tipo de discapacidad	Categórica	Cualitativa	Nominal	Politómico
TIPO_EMPLEO_NO_TIENE	Tipo de empleo - no tiene	Categórica	Cualitativa	Nominal	Politómico
TIPO_EMPLEO_PERMANENTE	Tipo de empleo - permanente	Categórica	Cualitativa	Nominal	Politómico
TIPO_EMPLEO_TEMPORAL	Tipo de empleo - temporal	Categórica	Cualitativa	Nominal	Politómico
TRABAJA	Condición actual de empleo	Categórica	Cualitativa	Nominal	Dicotómico
VECES_REPETIDO_ANO	Número de repitencias	Numérica	Cuantitativa	Razón	Discreta
VICT_DISCRI_LGBTI	Víctima de discriminación por pertenecer a grupo LGBTI	Categórica	Cualitativa	Nominal	Politómico
VICTIMA_AGRESIONES	Muestra si es víctima de agresiones	Categórica	Cualitativa	Nominal	Dicotómico
VIVE_SOLO	Vive solo	Categórica	Cualitativa	Nominal	Dicotómico
ZONA__ALU	Tipo de zona donde vive el alumno	Categórica	Cualitativa	Nominal	Dicotómico
ZONA_DE_ATENCIÓN	Zona de atención de la institución	Categórica	Cualitativa	Nominal	Dicotómico

Características de la data final

Después de revisar y comprender los datos, se tiene como resultado un total de 53 variables que le pueden aportar al estudio, entre ellas se encuentran 25 variables numéricas y 28 tipo texto; a continuación, se relacionan las variables finales:

Figura 8. *Listado de variables finales*

```
RangeIndex: 263075 entries, 0 to 263074
Data columns (total 53 columns):
 #   Column                                          Non-Null Count   Dtype
---  ------                                          --------------   -----
 0   VIVE_SOLO                                       263075 non-null  object
 1   EMBARAZO                                        71705 non-null   object
 2   TRABAJA                                         263075 non-null  object
 3   MAT_PAT_TEMPRANA_EDAD                           263075 non-null  object
 4   VICT_DISCRI_LGBTI                               246853 non-null  object
 5   VICTIMA_AGRESIONES                              260947 non-null  object
 6   NUMERO_PERSONAS_HOGAR                           262506 non-null  float64
 7   TENENCIA_VIVIENDA                               263075 non-null  object
 8   RET_EST_SIN_TERMINAR_ANO_ESC                    263075 non-null  object
 9   ABANDONOS_TEMPORALES                            71705 non-null   float64
 10  REPETIDA_ANO                                    263075 non-null  object
 11  VECES_REPETIDO_ANO                              228650 non-null  float64
 12  REPITIENDO_GRADO_ACTUAL                         71705 non-null   object
 13  ANTECED_DICIPL_VIDA_ACADEM                      71705 non-null   object
 14  ASIST_PROM_ANO_ANTERIOR                         71705 non-null   object
 15  DIF_APREN_DIAG_ESPEC                            263075 non-null  object
 16  NOM_ESTRAT_NOMBRE                               191988 non-null  object
 17  NOM_ESTRAT_CONTEO                               263075 non-null  int64
 18  ASISTE_REU_DIF_BOLETIN_ALGUNAS_VECES            228194 non-null  float64
 19  ASISTE_REU_DIF_BOLETIN_CASI_NUNCA               228194 non-null  float64
 20  ASISTE_REU_DIF_BOLETIN_CASI_SIEMPRE             228194 non-null  float64
 21  ASISTE_REU_DIF_BOLETIN_NUNCA                    228194 non-null  float64
 22  ASISTE_REU_DIF_BOLETIN_SIEMPRE                  228194 non-null  float64
 23  ASISTE ENTREGA INFORMES ALGUNAS VECES           228194 non-null  float64
 24  ASISTE_ENTREGA_INFORMES_CASI_NUNCA              228194 non-null  float64
 25  ASISTE_ENTREGA_INFORMES_CASI_SIEMPRE            228194 non-null  float64
 26  ASISTE_ENTREGA_INFORMES_NUNCA                   228194 non-null  float64
 27  ASISTE_ENTREGA_INFORMES_SIEMPRE                 228194 non-null  float64
 28  TIPO_EMPLEO_NO_TIENE                            228194 non-null  float64
 29  TIPO_EMPLEO_PERMANENTE                          228194 non-null  float64
 30  TIPO_EMPLEO_TEMPORAL                            228194 non-null  float64
 31  FRECUENCIA_CAMBIO_DOMIC_ENTRE_1_Y_2_VECES       228194 non-null  float64
 32  FRECUENCIA_CAMBIO_DOMIC_ENTRE_3_Y_4_VECES       228194 non-null  float64
 33  FRECUENCIA_CAMBIO_DOMIC_MAS_DE_4_VECES          228194 non-null  float64
 34  FRECUENCIA_CAMBIO_DOMIC_NO_HA_CAMBIADO          228194 non-null  float64
 35  APOYO ACADEMICO                                 259574 non-null  object
 36  PER_ID                                          263075 non-null  int64
 37  ESTRATO                                         263075 non-null  object
 38  SISBEN                                          263063 non-null  object
 39  GENERO                                          263075 non-null  object
 40  POBLACION_VICTIMA_CONFLICTO                     263075 non-null  object
 41  TIPO_DE_DISCAPACIDAD                            0 non-null       float64
 42  CAPACIDADES_EXCEPCIONALES                       0 non-null       float64
 43  ETNIA                                           263075 non-null  object
 44  GRADO                                           263075 non-null  object
 45  METODOLOGIA                                     263075 non-null  object
 46  REPITENTE                                       263075 non-null  object
 47  SITUACION_ACADEMICA_AÑO_ANTERIOR                256508 non-null  object
 48  CONDICION_DEL_ALUMNO_AL_FINALIZAR_EL_AÑO_ANTERIOR 260171 non-null  object
 49  ZONA__ALU                                       263075 non-null  object
 50  BENEF._MADRE_CABEZA_DE_FLIA                     263075 non-null  object
 51  ZONA_DE_ATENCIÓN                                263075 non-null  object
 52  EDAD                                            263075 non-null  int64
```

Mapeo indicador - atributo

A continuación, se presenta la relación de los atributos que contiene la base de datos utilizada y los indicadores o métricas planteadas en el presente estudio.

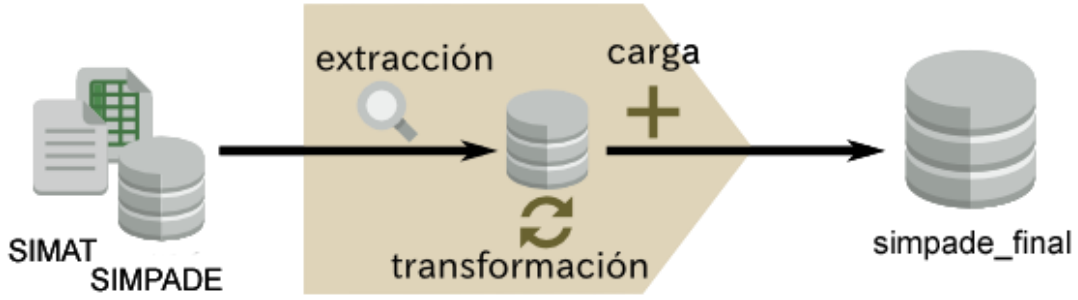
Tabla 4. *Relación atributos e indicadores*

Indicador o métrica	Cantidad de Atributos relacionados	Atributos
1. Total de posibles desertores generados por el modelo sobre el total de registros del modelo.	53	Total de atributos contenidos en la data Simpade_Final. Figura 8
2. Total posibles desertores por grado de escolaridad	2	CON_ALUMNO_AÑO_ANT
		GRADO
Total general de atributos	55	

9.1.3. Fase III. Análisis y Preparación de los Datos

Para desarrollar esta etapa del trabajo de grado, se realiza el proceso de ETL (Extracción, Transformación y Carga) con el fin de generar una sola fuente de datos; este proceso se basa en las siguientes fases:

Figura 9. *Proceso de ETL del trabajo*



Nota: Adaptado de *Proceso ETL*, por Dario Bernabeu, s.f., Hefesto

Extracción de datos

Se obtienen los datos en archivos planos en formato .csv directamente desde la Secretaría de Educación de Medellín, quienes suministraron la información a partir de dos fuentes, SIMAT y SIMPADE.

Transformación

Se realiza toda la preparación necesaria para obtener la base de datos final para la generación del modelo. Para ejecutar este proceso se utilizaron herramientas como Google Colab con Python y Excel, las cuales permitieron eliminar aquellos datos irrelevantes, desfasados, incorrectos y duplicados para el estudio; además, se pudo realizar la integración de ambas bases de datos por medio de la variable PER_ID (código único del estudiante), permitiendo así obtener desde la base de datos SIMAT, datos de las variables Tipo_De_Discapacidad y Capacidades_Excepcionales, las cuales no contenían registros en la base de datos de SIMPADE.

Actividad 5: Realizar la limpieza de los datos seleccionados

Informe de limpieza de datos

Se realiza proceso de limpieza para identificar datos erróneos y realizar la corrección adecuada para lograr datos de calidad para el estudio.

Se realiza corrección en Excel de los siguientes datos:

Se normaliza la variable embarazo con relación a la variable género; para todos los atributos HOMBRE se pone “NO APLICA” y para los atributos MUJERES en los campos vacíos se colocan “NO”. Todos los campos que en la variable REPETIDA_ANO dice “NO” se pone número de veces cero en la variable VECES_REPETIDA_ANO; de igual forma en los campos que dice “SI” y tienen cero o están vacías se reemplaza por la mediana; finalmente a los datos

por encima de 6 años de repitencia se aplica también la mediana por ser datos atípicos con la naturaleza de la variable. Se realiza transformación en la variable POBLACION_VICTIMA_CONFLICTO se pone "SI" en los campos que están caracterizados como víctimas y "NO" en los que no aplica caracterización; de igual forma en la variable ETNIA se pone "SI" en los campos que están caracterizados pertenecientes a una etnia y "NO" en los que no aplica. Se corrige la variable SISBEN sustituyendo los ceros por "NO APLICA".

Se hace uso de la biblioteca Pandas para limpieza con lenguaje Python de los siguientes datos:

Se eliminan las variables NUMERO_PERSONAS_HOGAR, ASISTE_ENTREGA_INFORMES_ALGUNAS_VECES, ASISTE_ENTREGA_INFORMES_CASI_NUNCA, ASISTE_ENTREGA_INFORMES_CASI_SIEMPRE, ASISTE_ENTREGA_INFORMES_NUNCA y ASISTE_ENTREGA_INFORMES_SIEMPRE por ser la mayoría de sus datos atípicos. Se eliminan las variables ABANDONOS_TEMPORALES y APOYO ACADEMICO porque tienen gran cantidad de datos faltantes con relación al total de registros. Las variables REPITIENDO_GRADO_ACTUAL y REPITENTE se eliminan, dado que la información que contienen no es coherente con la información que suministra la variable REPETIDA_ANO y esta última proporciona más información útil para el estudio. Se sustituyen los datos nulos de las siguientes variables por "NO APLICA": ASIST_PROM_ANO_ANTERIOR, ANTECED_DICIPL_VIDA_ACADEM, VICT_DISCRI_LGBTI, VICTIMA_AGRESIONES, SISBEN, SIT_ACADEMICA_AÑO_ANT y CON_ALUMNO_AÑO_ANT.

No se cuenta dentro de la base de datos con registros duplicados

```
#Eliminar Registros Duplicados
simpade_final.loc[simpade_final.duplicated()]
```

PER_ID	VIVE_SOLO	EMBARAZO	TRABAJA	MAT_PAT_TEMPRANA_EDAD	VICT_DISCRI_LGBTI	VICTIMA_AGRESIONES

0 rows × 43 columns

Actividad 6: Integrar los datos seleccionados como necesarios de las bases conseguidas desde los diferentes sistemas de información.

Integración de datos para realizar el estudio desde las fuentes SIMAT y SIMPADE

Las variables TIPO_DE_DISCAPACIDAD y CAPACIDADES_EXCEPCIONALES no contienen registros en la base SIMPADE, por lo tanto, se integran los datos con la base SIMAT, posteriormente se elimina la variable PER_ID, que ya no es necesaria dentro del estudio.

Carga de datos

Por último, en esta fase se procede a cargar en la herramienta Google Colab la base de datos resultante de la fase anterior, y con ésta se procederá a generar los modelos de Machine Learning para obtener el modelo óptimo de acuerdo con los objetivos del estudio.

Visualizaciones de los datos

Variable objetivo: el atributo condición alumno año anterior (CON_ALUMNO_AÑO_ANT) cuenta con pocos registros tipificados como deserción con relación al total de la base de datos como se puede observar en la Figura 10, razón por la cual es necesario realizar balanceo de esta variable para evitar sesgos en el estudio.

Figura 10. *Visualización variable dependiente - Condición del alumno año anterior*

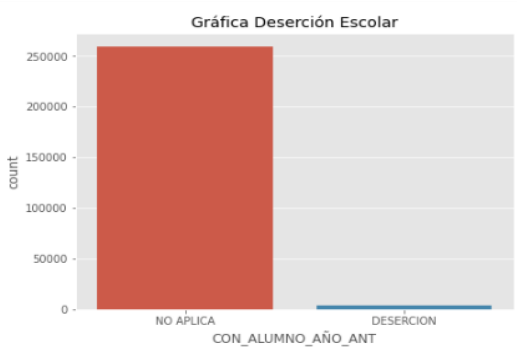
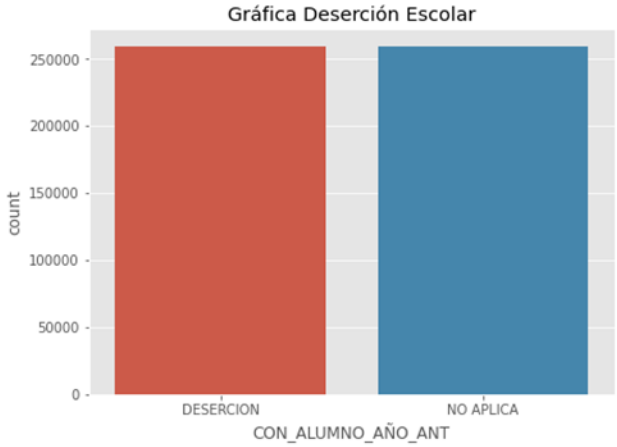


Figura 11. *Visualizaciones de las variables categóricas que contiene la base de datos*

Balanceo de la variable objetivo

Debido a que existe un desbalance amplio de la variable objetivo, se realiza balanceo de los datos usando la técnica oversampling (creando nuevos registros en la clase minoritaria); dado que la variable objetivo cuenta con 3.664 registros tipificados como “DESERCION” sobre el total que son 263.075, lo que equivale solo al 1.39% del total de los datos. Se modifica la distribución original de los datos generando registros tipificados con el minoritario “DESERCIÓN” e igualando los registros al total tipificados como “NO APLICA”, como se muestra en la Figura 12.

Figura 12. *Balanceo variable dependiente*




```
#Contar datos
d_simpade.groupby('CON_ALUMNO_AÑO_ANT').CON_ALUMNO_AÑO_ANT.count().sort_values(ascending=False)
```

```
CON_ALUMNO_AÑO_ANT
DESERCION    259411
NO APLICA    259411
Name: CON_ALUMNO_AÑO_ANT, dtype: int64
```

9.2. Desarrollo del Objetivo Específico 2: Aplicar Diferentes Técnicas Supervisadas de clasificación sobre los Datos Preparados para Generar el Modelo Óptimo

El desarrollo de este objetivo comprende la fase IV de la metodología CRISP DM, utilizada en el desarrollo del trabajo. Esta fase se divide en dos actividades.

9.2.1. Fase IV Modelado

Actividad 7: Seleccionar las técnicas de modelado a aplicar, acorde con los datos y objetivos.

De acuerdo con el objetivo del estudio y los datos que se tienen donde se conoce la variable objetivo; se identificó que las técnicas que pueden generar los resultados deseados son: Regresión Logística, Árbol de Decisión y Random Forest; para la aplicación de estas técnicas se realiza inicialmente transformación de las variables categóricas a variables dummies como se observa en la Figura 13.

Figura 13. *Generación de Variables Dummies*

```
d_simpade.head()
```

	VECES_REPETIDO_ANO	NOM_ESTRAT_CONTEO	ASISTE_REU_DIF_BOLETIN_ALGUNAS_VECES	ASISTE_REU_DIF_BOLETIN_CASI_NUNCA
196697	1	0	0	0
220444	0	0	0	0
60914	0	2	0	0
3159	1	1	0	0
48256	1	2	0	0

5 rows × 77 columns

División del conjunto de datos para entrenamiento y evaluación

Los datos finales después del procesamiento y balanceo se dividen en dos conjuntos, El primer conjunto de datos, es el de entrenamiento (train), el cual se le asigna un valor del 80% de los registros y el segundo, es el conjunto de datos de evaluación (test), al cual se asigna el

20% restante de los registros. Lo anterior, es con el fin de poder evaluar el rendimiento del modelo diseñado con datos nuevos que no se utilizaron en el entrenamiento.

Escalar variables

Se llevan las variables numéricas a una misma escala para conocer exactamente la distancia entre ellas y que no se presenten variables que por tener rangos muy diferentes influyan más que otras en el cálculo de distancia con relación a la variable dependiente.

Figura 14. *Muestra variables escaladas*

	VECES_REPETIDO_ANO	NOM_ESTRAT_CONTEO	ASISTE_REU_DIF_BOLETIN_ALGUNAS_VECES	ASISTE_REU_DIF_BOLETIN_CASI_NUNCA
211588	0.000000	0.000000	0.0	0.0
6655	0.166667	0.166667	0.0	0.0
76994	0.166667	0.333333	0.0	0.0
82690	0.166667	0.333333	0.0	0.0
221002	0.000000	0.000000	0.0	0.0

Actividad 8: Construir los modelos a partir de la aplicación de las técnicas seleccionadas.

Creación y evaluación de los modelos

Se utiliza el conjunto de entrenamiento en las técnicas de Regresión Logística, Árbol de Decisión y Random Forest y se establecen los parámetros de calibración y construcción de cada uno de los modelos para la predicción de la deserción escolar y la identificación de las variables con mayor peso con relación a la variable dependiente. Posteriormente se validan los 3 modelos con el grupo de datos de evaluación, para determinar su desempeño basado en la predicción de los datos restantes; los tres modelos se evalúan en términos de la exactitud, exhaustividad/sensibilidad (recall), precisión y el valor F1, obteniendo para todas ellas valores por encima del 99% de acuerdo con las Figuras 17, 19 y 21; donde se muestra la evaluación realizada a cada modelo y las métricas generadas; sin embargo con el modelo de Regresión Logística se observan resultados más significativos para el estudio, dado que arroja los

coeficientes y estadísticos de cada variable, es decir la importancia de las variables dependientes respecto a la variable objetivo de este trabajo, lo que arroja información importante para el logro de los objetivos propuestos.

En las matrices de confusión desarrolladas los valores de la diagonal principal indican que la mayoría de las clasificaciones de la condición del estudiante están correctamente clasificadas, tanto los que, sí son desertores, como los que no desertaron; mientras que la otra diagonal, que representa los casos en los que el modelo se ha equivocado (falsos desertores y falsos no desertores) tiene valores muy pequeños, que no afectan en gran medida la precisión de los modelos.

Para la creación del modelo de árbol de decisión se utiliza la librería scikit-learn con este árbol de clasificación se busca hallar las variables que más influyen en la deserción escolar. Después de realizar varias pruebas con diferentes parámetros, se opta por desarrollar el modelo aplicando los siguientes parámetros: max_depth = 5 (profundidad máxima del árbol) y random_state = 123 (controla la aleatoriedad del estimador); posterior a la creación del modelo, se procede a realizar la evaluación del modelo sobre los datos de test para determinar la precisión (accuracy), la cual arrojó un valor de 99.1%, cómo se puede observar en la Figura 15.

Figura 15. *Creación y evaluación del modelo Árbol de decisión*

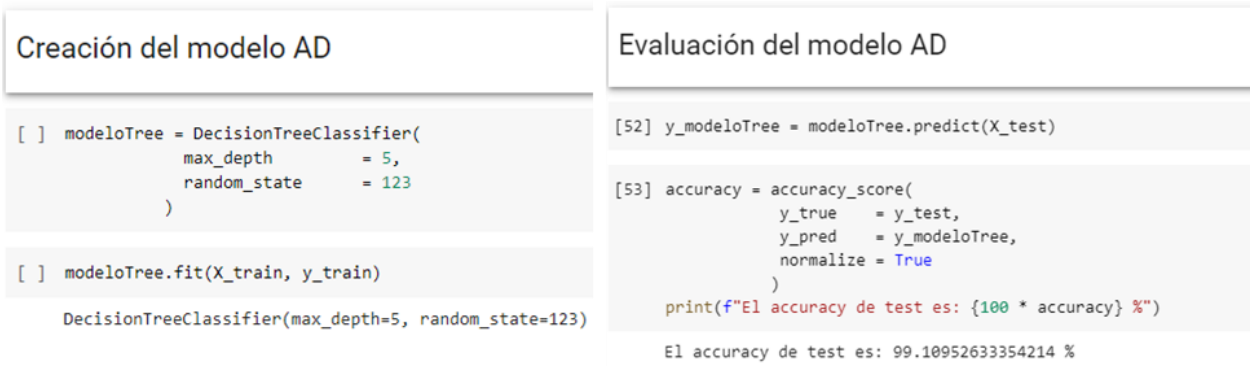


Figura 16. *Estructura del árbol creado*

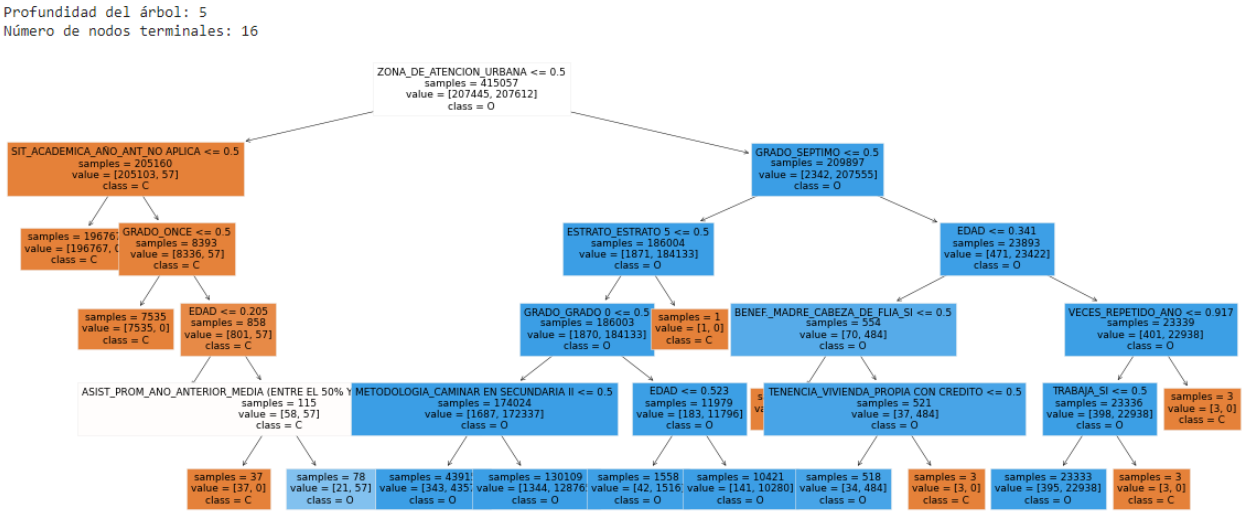
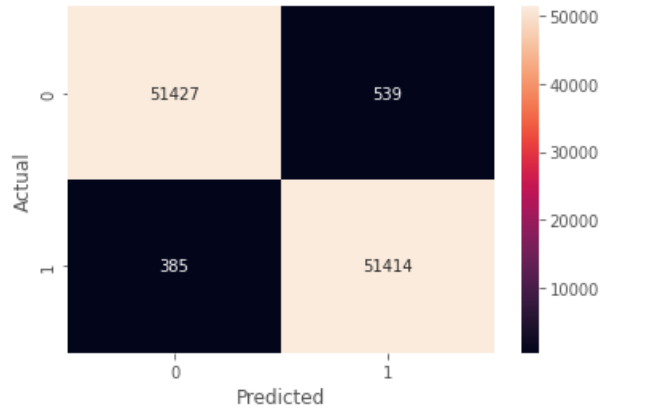


Figura 17. *Matriz de confusión del modelo de Árbol de Decisión*



	precision	recall	f1-score	support
0	0.99	0.99	0.99	51966
1	0.99	0.99	0.99	51799
accuracy			0.99	103765
macro avg	0.99	0.99	0.99	103765
weighted avg	0.99	0.99	0.99	103765

Continuando con la creación de los modelos, se creó el modelo de Random Forest, para el cual se utilizó a través de la librería sklearn.ensemble RandomForestClassifier, para generar un modelo de clasificación. Luego de realizar varias pruebas con diferentes parámetros, se optó por desarrollar el modelo aplicando los siguientes parámetros: n_stimators = 40 (Número

de árboles en el bosque), max_depth = 5 (profundidad máxima del árbol) y random_state = 123 (controla la aleatoriedad del estimador). Posterior a la creación del modelo y cómo se puede observar en la Figura 18, se procede a realizar la evaluación del modelo sobre los datos de test para estimar la precisión, por lo que se valida el error cuadrático medio (MSE) y las puntuaciones de R-cuadrado, los cuales arrojaron valores muy buenos de 0.09 y 0.96 respectivamente.

Figura 18. *Creación y evaluación del modelo de Random Forest*

Creación del modelo RF

```
[56] modelRanForest = RandomForestClassifier(
         n_estimators=40, max_depth=8,
         random_state=0).fit(X_train, y_train)

[57] y_RanForest= modelRanForest.predict(X_test)
```

Evaluación del modelo RF

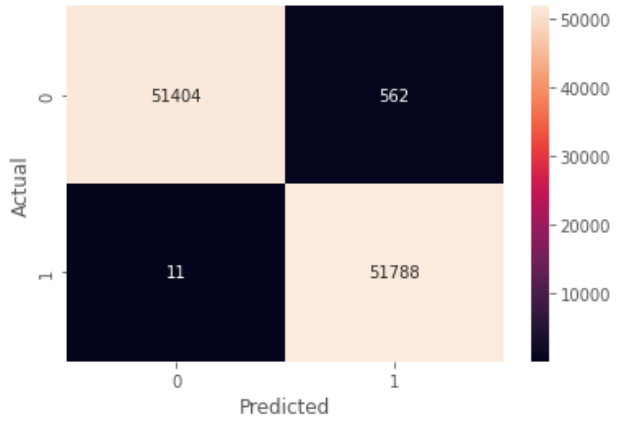
```
[ ] prediccionesRF = modeloRF.predict(X = X_test)

    rmse = mean_squared_error(y_true  = y_test, y_pred  = prediccionesRF, squared = False)
    print(f"El error (rmse) de test es: {rmse}")

    r2 = r2_score(y_test, prediccionesRF)
    print(f"El error (r2) de test es: {r2}")

    El error (rmse) de test es: 0.09428424956514114
    El error (r2) de test es: 0.9644353018492698
```

Figura 19. *Matriz de confusión del modelo Random Forest*



	precision	recall	f1-score	support
0	1.00	0.99	0.99	51966
1	0.99	1.00	0.99	51799
accuracy			0.99	103765
macro avg	0.99	0.99	0.99	103765
weighted avg	0.99	0.99	0.99	103765

Finalmente, se crea un modelo de Regresión Logística, para el cual se utilizaron las librerías de Scikit-Learn y Statsmodels. Después de realizar varias pruebas aplicando

diferentes parámetros al modelo, se optó por desarrollar el modelo con el Statsmodels, dado que entrega los coeficientes y estadísticos de cada una de las variables, necesarios para el análisis y verificación de cumplimiento de las condiciones sobre las que se basa este tipo de modelos, las métricas analizadas en este modelo fueron el accuracy como se muestra en la Figura 20, del 99,45%.

Figura 20. *Creación y evaluación del modelo de Regresión Logística*

Creación del modelo RL

```
[60] X_train = sm.add_constant(X_train)
     modeloLR = sm.Logit(y_train, X_train)
     modeloLR = modeloLR.fit_regularized()
     print(modeloLR.summary())
```

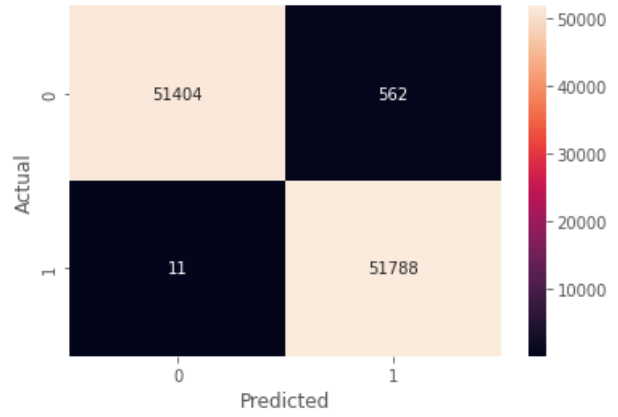
```
Optimization terminated successfully    (Exit mode 0)
            Current function value: 0.031112843872888766
            Iterations: 465
            Function evaluations: 465
            Gradient evaluations: 465
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:               415057
Model:                          Logit   Df Residuals:                   414978
Method:                           MLE   Df Model:                           78
Date:                Thu, 24 Nov 2022   Pseudo R-squ.:                  0.9551
Time:                        04:17:12   Log-Likelihood:                -12914.
converged:                       True   LL-Null:                   -2.8770e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
```

Evaluación del modelo RL

```
[61] X_test = sm.add_constant(X_test)
     predicciones = modeloLR.predict(X_test)
     clasificacion = np.where(predicciones<0.5, 0, 1)
     accuracy = accuracy_score(
                y_true    = y_test,
                y_pred    = clasificacion,
                normalize = True
               )
     print("")
     print(f"El accuracy de test es: {100*accuracy}%")
```

```
El accuracy de test es: 99.44779068086542%
```

Figura 21. *Matriz de confusión del modelo Regresión Logística*



	precision	recall	f1-score	support
0	1.00	0.99	0.99	51966
1	0.99	1.00	0.99	51799
accuracy			0.99	103765
macro avg	0.99	0.99	0.99	103765
weighted avg	0.99	0.99	0.99	103765

9.3. Desarrollo del Objetivo Específico 3: Evaluar el Modelo Generado para la Identificación de las Variables Predictoras que tienen Mayor Peso o Influencia en la Deserción Estudiantil.

El desarrollo de este objetivo comprende la fase V de la metodología CRISP DM, utilizada en el desarrollo del trabajo. Esta fase se divide en tres actividades.

Actividad 9: Utilizar el conjunto de datos de test sobre los modelos generados para su evaluación.

9.3.1. Fase V – Evaluación

Descripción del modelo aprobado

Se creó un modelo de Regresión Logística con un Pseudo R2 de 0.9552, es decir con un ajuste casi perfecto, lo que lo hace un modelo confiable para generar información en términos explicativos y predictivos, logrando conocer las variables estadísticamente significativas para el modelo y la probabilidad de que los alumnos activos deserten de las instituciones educativas.

Métricas modelo Regresión Logística

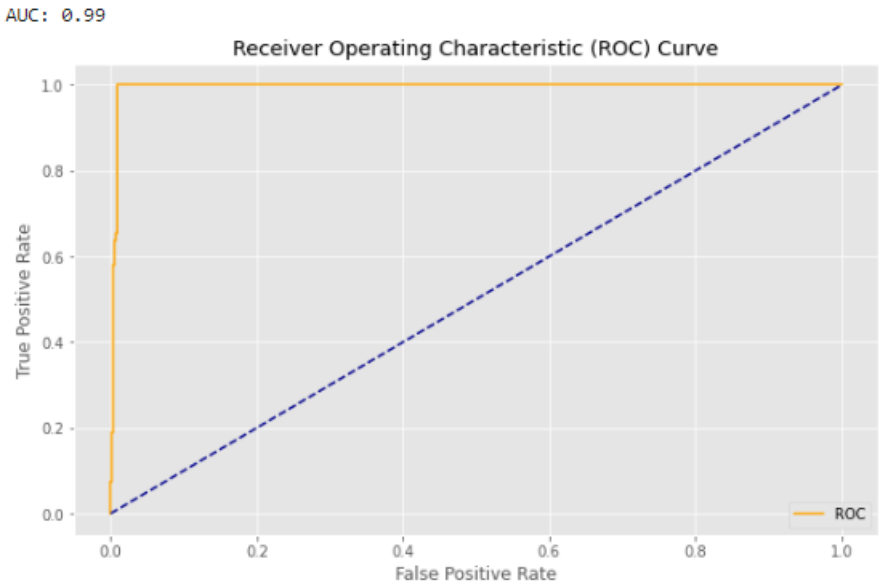
La Figura 21 muestra el rendimiento del modelo de regresión logística con el conjunto de datos de prueba; la mayoría de las clasificaciones están correctamente clasificadas, la exactitud se cuantifica las predicciones correctas frente al total de casos del modelo, es decir (TP+TN)/(TP+FP+FN+TN); para este caso es (51788+51404)/(51788+562+11+51404) para lo cual obtenemos una exactitud del 99,45%. La sensibilidad (o tasa de verdaderos positivos) es la proporción de los que se pronosticaron como desertores entre los que realmente abandonaron la escuela TP/(TP+FN), 51788/ (51788 + 11) la sensibilidad de este modelo da

99,98%. La especificidad (o tasa negativa verdadera) es la proporción de los que se pronostica que no abandonarán entre los que realmente permanecerán en la escuela. La especificidad es TN/(TN+FP) = 51404/ (51404 +562) = 98,92%.

Curva ROC

Como se puede observar en la Figura 22, la curva ROC es una representación gráfica de la sensibilidad frente a la especificidad e indica que este modelo es bueno para distinguir las clases dadas, de acuerdo con la probabilidad predicha, lo que indica el buen desempeño del modelo. Esta observación se cuantifica por el área bajo la curva ROC (es decir, AUC), que es 0,99 para el modelo desarrollado.

Figura 22. *Curva ROC del modelo*



Actividad 10: Realizar ajustes en el proceso, en caso de ser necesario

No fue necesario realizar ajustes en los modelos creados, dado que, con las métricas utilizadas para su creación, se puedo observar en los resultados obtenidos que los modelos estaban cumpliendo con el alcance del objetivo de este trabajo, ya que la precisión de todos los modelos estuvo por encima del 99%.

Actividad 11: Realizar revisión de los resultados obtenidos con el modelo construido respecto con los propósitos de analítica propuestos.

Informe de revisión (resultados vs propósitos de analítica)

Uno de los propósitos de análisis de este trabajo es el identificar las variables que tienen mayor incidencia en la deserción estudiantil; este modelo de Regresión Logística entregó 14 variables estadísticamente significativas (p-value por debajo de 0.05), para la predicción de la deserción escolar lo que corresponde a un 18% del total de las variables que contiene la data y donde cuyos intervalos de confianza no incluyen el 0, lo que corrobora su significancia. Dentro de las variables significativas tenemos la edad, así que, ¿Cuál es la probabilidad de que una persona deserte con relación a la edad?, para esta variable se tiene un coeficiente positivo de 5.1933 lo que significa que entre más años tenga un estudiante, más alta es la probabilidad de abandonar sus estudios; así mismo sucede cuando un alumno pertenece al grado primero o tercero, es beneficiario de programas para madres cabeza de familia o si no estudió en la vigencia académica del año anterior, tiene mayor probabilidad de desertar; dado que estas variables también presentan coeficientes positivos dentro del modelo generado; caso contrario pasa con las variables para las que se obtuvo coeficientes negativos, éstas indican que si se pertenece a esa clase, se tiene menor probabilidad de abandonar los estudios; es así como tenemos que los estudiantes que no han sido víctimas de discriminación LGTBI tienen menos probabilidad de desertar, al igual que los que tienen vivienda obtenida mediante crédito, los que son de grado noveno, once y quinto, y los que están en las metodologías caminar en secundaria I y II. Un resultado realmente significativo es que entre más un alumno acceda a las estrategias de permanencia escolar que brinda la institución educativa menor es la probabilidad de desertar, para esta variable se obtuvo un coeficiente de -0.6038.

Figura 23. *Variables con mayor influencia en la deserción*

	coef	std err	z	P>|z|	[0.025	0.975]
NOM_ESTRAT_CONTEO	-0.6038	0.210	-2.877	0.004	-1.015	-0.193
EDAD	5.1933	0.446	11.638	0.000	4.319	6.068
VICT_DISCRI_LGBTI_NO APLICA	-0.1638	0.042	-3.857	0.000	-0.247	-0.081
TENENCIA_VIVIENDA_PROPIA CON CREDITO	-0.3690	0.122	-3.024	0.002	-0.608	-0.130
GRADO_NOVENO	-1.5177	0.097	-15.596	0.000	-1.708	-1.327
GRADO_ONCE	-1.5508	0.186	-8.316	0.000	-1.916	-1.185
GRADO_PRIMERO	1.1850	0.117	10.147	0.000	0.956	1.414
GRADO_QUINTO	-0.3498	0.083	-4.236	0.000	-0.512	-0.188
GRADO_TERCERO	0.3558	0.098	3.640	0.000	0.164	0.547
METODOLOGIA_CAMINAR EN SECUNDARIA I	-0.3465	0.103	-3.355	0.001	-0.549	-0.144
METODOLOGIA_CAMINAR EN SECUNDARIA II	-0.4286	0.150	-2.855	0.004	-0.723	-0.134
BENEF._MADRE_CABEZA_DE_FLIA_SI	0.2922	0.098	2.972	0.003	0.100	0.485
SIT_ACADEMICA_AÑO_ANT_NO ESTUDIO EN LA VIGENCIA ANTERIOR	9.4409	3.944	2.394	0.017	1.711	17.171

El segundo propósito de analítica en este trabajo es generar un modelo que permita predecir los posibles desertores de las instituciones educativas; el modelo se generó con la técnica de Regresión Logística y con la evaluación del modelo se obtuvo resultados muy positivos, con métricas casi perfectas, lo que representa que los datos que se pueden obtener del modelo pueden generar información importante para la toma de decisiones. Adicionalmente a la evaluación, donde se obtuvo predicciones con alta precisión; el modelo se aplicó a una nueva data, clasificando a 7.190 estudiantes como posibles desertores, lo que equivale al 2.71% del total de la matrícula 2022, con corte a 31 de octubre, de las instituciones educativas de Medellín de estudiantes en edad regular.

Tabla 5. *Clasificaciones del modelo*

ATRIBUTO	**CON_ALUMNO_AÑO_ANT**
DESERCION	7139
NO_APLICA	258321
Total de registros	**265460**

9.4. Desarrollo del Objetivo Específico 4: Presentar los Resultados Obtenidos con el Modelo Desarrollado sobre la Predicción de Posibles Desertores de acuerdo con las Variables Obtenidas que tienen Mayor Peso.

El desarrollo de este objetivo comprende la fase Vl de la metodología CRISP DM, utilizada en el desarrollo del trabajo. Esta fase se divide en dos actividades.

9.4.1. Fase Vl. Despliegue

Para el desarrollo de esta fase, se realizará la entrega del modelo construido de Regresión Logística al Observatorio para la Calidad Educativa de Medellín (OCEM), además del informe generado de las variables con mayor incidencia en la deserción y la base de datos en formato csv con predicciones de posibles desertores después de realizar pruebas del modelo con la base de matrícula actual. Como recomendación y para que esta herramienta alcance el valor deseado para el OCEM, se recomienda al OCEM, tener en cuenta los resultados expresados en los entregables, además de compartirlos con el área encargada de la permanencia escolar de los estudiantes del Distrito de Medellín y las demás dependencias y entidades a las que pueda contribuir este estudio, con el fin de que sirva de insumo para apoyarse en la asignación de estrategias de permanencia de los estudiantes.

Actividad 12: aplicar el modelo construido en una base de datos de matrícula vigente de instituciones educativas de la ciudad de Medellín

Actividad 13: Generar un informe con los resultados obtenidos de acuerdo con los objetivos del proyecto.

Informe de resultados del estudio con Machine Learning sobre la deserción escolar para Secretaría de Educación de Medellín

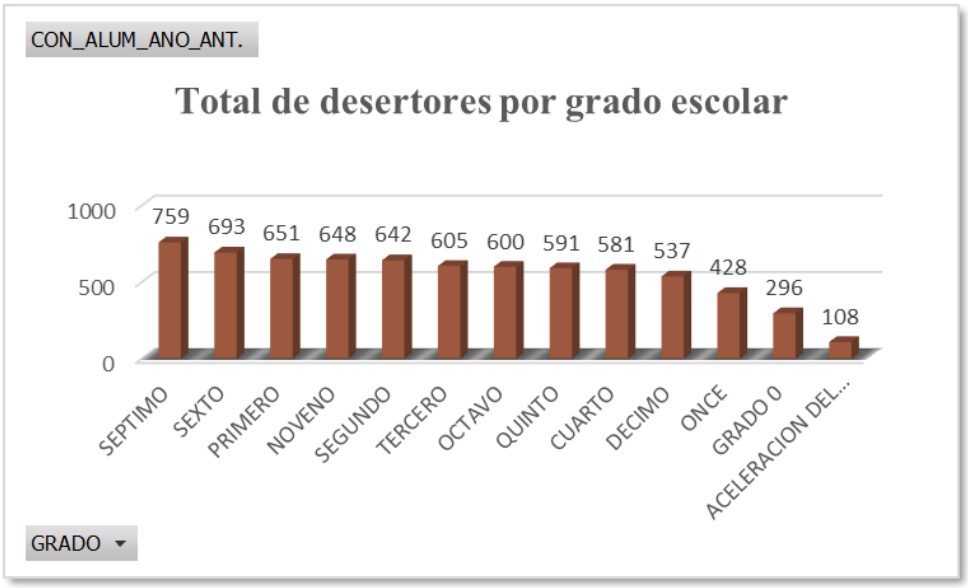
Para el desarrollo de este estudio se utilizaron 259.411 registros de estudiantes matriculados para el 2020 en instituciones educativas oficiales de Medellín en edad regular con base en reportes de los sistemas SIMAT y SIMPADE; con estos datos se creó un modelo predictivo bajo la técnica de Regresión Logística con una precisión del 99%, a través del cual fue posible obtener información importante que pueda contribuir a la evaluación, seguimiento e implementación de estrategias de permanencia. Los objetivos propuestos para el desarrollo de este estudio fueron alcanzados con el modelo diseñado, dado que se obtuvo información de cuáles pueden ser las variables que más peso tienen o más influyen en la deserción de los estudiantes y se obtuvo información de qué estudiantes podrían ser posibles desertores. La información generada puede facilitar al área encargada la implementación de estrategias de permanencia escolar para la toma de decisiones en función de cómo abordar la deserción escolar desde una perspectiva de asignar las estrategias correctas a un estudiante para que no deserte del sistema educativo.

De acuerdo con los resultados obtenidos con el modelo creado; como se mencionó anteriormente, se pudo identificar variables directamente relacionadas con la condición de desertar, es decir si esa variable aumenta o es positiva, la posibilidad de abandonar los estudios también aumenta, estas variables fueron: edad, pertenencia al grado primero o tercero, ser beneficiario de programas para madres cabeza de familia o no estudiar en la

vigencia académica del año anterior; por ejemplo, cada que un estudiante aumenta su edad, se aumenta la probabilidad de abandonar sus estudios. Así también se obtuvo variables indirectamente relacionadas, donde al aumentar o ser positiva esta variable disminuye la posibilidad de deserción, tales variables son: no ser víctimas de discriminación LGTBI, tener vivienda obtenida mediante crédito, ser de grado noveno, once y quinto, estar en las metodologías caminar en secundaria I y II y acceder a las estrategias de permanencia escolar; un ejemplo de esto, es que entre más estrategias de permanencia tenga un estudiante, menor es el riesgo de desertar.

Adicionalmente, para obtener resultados de los estudiantes con posibilidad de deserción, se aplicó el modelo creado a la base de matrícula 2022 con corte a 31 de octubre, con el fin de predecir cuáles estudiantes de los que se encuentran actualmente matriculados tienen riesgo de desertar, obteniendo como resultado 7.139 posibles desertores (se entrega a la Secretaría archivo csv con los resultados completos de las predicciones), estos estudiantes tienen alguno de los atributos de las variables identificadas como de mayor peso para el abandono escolar, además con base en los registros obtenidos, se pueden realizar comparación de los posibles desertores a las características comunes que poseen; por ejemplo, se puede clasificar a los desertores por grado escolar como se ve en la Figura 23 dando como resultado que los estudiantes con mayor riesgo de deserción pertenecen al grado séptimo; sin embargo no se presentan variaciones representativas en la mayoría de los casos.

Figura 24. *Posibles desertores clasificados por grado escolar*



Recomendaciones:

Buscar estrategias que sirvan para mejorar la recolección de datos a través de los sistemas de información SIMAT y SIMPADE, dado que estos sistemas fueron fundamentales para el desarrollo de este estudio y de la fase de despliegue en la aplicación del modelo a la data actual, no obstante, se encontró mucho registro con datos faltantes y en ceros; se debe fortalecer en las instituciones educativas que son las responsables de esta recolección de información, para obtener mejores datos que generen la toma de decisiones más certeras.

Se realizará entrega del modelo desarrollado en un archivo ipynb tipo Python, el cual puede ser abierto en una herramienta de trabajo como Colab, herramienta utilizada para el desarrollo de este trabajo de grado. Se recomienda que la persona que aplique el modelo tenga conocimiento en la herramienta, que le permita interactuar con el modelo y explotar la información

Se recomienda realizar una interfaz de ejecución del modelo entregado, a través del área de tecnología; dado que, para la aplicación, la persona que lo vaya a realizar requiere tener conocimientos básicos en Python y esto puede llevar a que no sea fácil su ejecución.

El modelo entregado da respuesta a los objetivos planteados en este proyecto, por lo cual podrá ser un insumo importante para el área de permanencia de la Secretaría de Educación de Medellín, ya que le permitirá conocer cuáles pueden ser las variables que más están influyendo en la deserción escolar, de igual forma conocer cuáles estudiantes pueden llegar a ser posibles desertores.

.

10. Discusión

Abandonar el sector educativo para los estudiantes se puede transformar en costos sociales adicionales para la sociedad. A pesar de las diferentes estrategias implementadas por la Secretaría de Educación de Medellín enfocadas tanto en ayudar a los niños, niñas y jóvenes no escolarizados como las implementadas a los abandonos, se debe crear un mayor enfoque en identificar los posibles desertores y cómo prevenir que los estudiantes salgan de las instituciones educativas. Con el modelo predictivo Machine Learning desarrollado se tiene la posibilidad de identificar los estudiantes en riesgo de desertar para así poder implementar estrategias que puedan ayudarlos. Para este estudio se exploraron tres modelos supervisados como lo fueron los Árboles de Decisión, Random Forest y Regresión Logística, encontrando este último como el modelo óptimo para predecir la deserción escolar y hallar los pesos de las variables que más influyen en esta problemática.

Las métricas de desempeño del modelo de regresión logística mostraron una excelente precisión obteniendo un valor del 99%; una exactitud de las predicciones correctas frente al total del 99,45%; una sensibilidad del 99,98% y una especificidad del 98,92%. Como comparación al estudio realizado, se puede tomar como referencia lo realizado en Corea, donde diseñaron un modelo predictivo con la técnica de Random Forest para generar alertas tempranas de abandono para los estudiantes de secundaria, con él se pudo obtener una métrica de precisión del 95%. (Chung & Lee. 2019), por lo cual, se puede decir que el modelo desarrollado en este estudio tiene muy buenos valores para realizar la predicción de la deserción escolar y al igual que el aplicado en Corea, generar alertas tempranas.

De acuerdo con la revisión de literatura realizada para elaborar este trabajo, se puede concluir que hay muchos factores que están relacionados con el abandono de los estudiantes;

sin embargo, varios de los factores discutidos en la revisión presentan muchos datos faltantes en las bases suministradas por la Secretaría de Educación, por lo cual un modelo predictivo que utilice una data con variables que tengan los registros más completos, factores que no se tienen en el modelo actual, podría mejorar el rendimiento de la predicción para futuros estudios.

11. Conclusiones

Del primer objetivo específico se puede concluir que:

En el procesamiento de los datos se identificó que las bases generadas a través de los sistemas SIMAT y SIMPADE entregadas para el desarrollo de este estudio contienen muchos datos nulos, por lo cual cabe resaltar la importancia del adecuado registro de la información por parte de las instituciones educativas con el fin de generar soluciones óptimas para el desarrollo de estrategias efectivas.

Con las técnicas de limpieza de datos aplicadas, se logró corregir y eliminar registros inexactos e irrelevantes para tener una fuente de datos con calidad, lo que permitió mejorar la precisión del modelo desarrollado.

Del segundo objetivo específico se puede concluir que:

A partir de las bases de datos suministradas por la Alcaldía de Medellín se crearon y evaluaron tres modelos diferentes a través de técnicas de Machine Learning, para la predicción de posibles desertores de instituciones educativas de Medellín en edad regular; todos los modelos arrojaron métricas de exactitud, exhaustividad/sensibilidad (recall), precisión y F1 por encima de 99%, lo que indica que las técnicas manejadas en este estudio pueden ser utilizadas en futuras investigaciones teniendo en cuenta el objeto de estudio.

Para este caso en particular donde uno de los propósitos era el encontrar las variables con mayor incidencia en la deserción escolar, se optó por trabajar con el modelo generado por la técnica de Regresión Logística (LR), pues éste entrega valores que permiten saber qué variables son estadísticamente significativas con relación a la variable condición del alumno el año anterior, la cual incluye el atributo estudiado de la deserción.

Del tercer objetivo específico se puede concluir que:

El modelo LR presentó métricas generales del 99% para la predicción de deserción escolar; aunque son leves las diferencias entre las predicciones para cada clase; la clase 0 (no desertor) tiene métrica más alta para la precisión que la sensibilidad, por lo tanto, no es perfecta la detección de la clase, pero cuando lo hace es altamente confiable; por el contrario la clase 1 (desertor) tiene menor precisión y mayor sensibilidad, lo que la hace detectar perfectamente la clase, pero también incluye muestras de la clase 0.

La deserción escolar es un problema que involucra múltiples factores, el modelo identificó que 14 variables tienen gran incidencia en la deserción escolar y por lo tanto se recomienda enfocar en ellas la generación y fortalecimiento de las estrategias que la Alcaldía ofrece a los estudiantes con el fin de aumentar la permanencia estudiantil y lograr un mejor aprovechamiento de los recursos destinados para este fin.

Del cuarto objetivo específico se puede concluir que:

La aplicación del modelo en una base de matrícula actual de 265.460 estudiantes, clasificó 7.139 como posibles desertores; siendo el grado séptimo el que presenta mayor número con 759; en general no son muy significativas las variaciones grado a grado, con excepción de los dos primeros grados, aceleración y cero, y el último, once, como se puede ver en la figura 23; es importante evaluar los atributos que tienen estos desertores, con el fin de que estos datos puedan contribuir a la asignación efectiva de beneficios para la permanencia del estudiante en la institución educativa frente a las variables identificadas para cada uno de ellos.

La aplicación de minería de datos a nivel educativo implementada por medio de modelos predictivos de aprendizaje ofrece una solución comprensible y de precisión, dado que

al identificar los factores que inciden en la deserción escolar, se pueden establecer y/o fortalecer las estrategias de permanencia escolar que ayuden a garantizar el no abandono de los estudiantes del sector educativo.

12. Diagrama de la Estructura del Trabajo

13. Referencias

Acuerdo 2 de 2020. [Concejo de Medellín]. Por el cual se adopta el Plan de Desarrollo Medellín Futuro 2020 – 2023. 31 de mayo de 2020

Adelman, M., Haimovich, F., Ham, A., & Vázquez, E. (2018). Predicting school dropout with administrative data: new evidence from Guatemala and Honduras. *Education Economics*, *26*(4), 356–372. https://doi.org/10.1080/09645292.2018.1433127

Agrega. (s.f.). Python en la nube con Google Colab. https://www.agrega.com/blog/general/python-en-la-nube-con-google-colab/

Aguirre, C. E., & Perez, J. C. (2020). Predictive data analysis techniques applied to dropping out of university studies. *Proceedings - 2020 46th Latin American Computing Conference, CLEI 2020*, 512–521. https://doi.org/10.1109/CLEI52000.2020.00066

Álvarez, D. (2021). Metodología CRISP-DM. https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/

Ansari, A., & Gottfried, M. A. (2018). The Benefits of Center-Based Care and Full-Day Kindergarten for School Attendance in the Early Grades. *Child and Youth Care Forum*, *47*(5), 701–724. https://doi.org/10.1007/s10566-018-9453-2

Ansari, A., & Gottfried, M. A. (2021). The Grade-Level and Cumulative Outcomes of Absenteeism. Child Development, 92(4), e548–e564. https://doi.org/10.1111/cdev.13555

Apaza, L. A. v, Huamani, J. A. R., Bernedo, J. O. A., & Chauca, A. G. Z. (2021). A proposal of Machine Learning model to improve learning process and reduce dropout rate at

technical training institutes. *Iberian Conference on Information Systems and Technologies, CISTI*. https://doi.org/10.23919/CISTI52073.2021.9476370

AprendeIA. (2019). Definición de Bosques Aleatorios. https://aprendeia.com/bosques-aleatorios-regresion-teoria-machine-learning/

AprendeIA. (2019). Definición de Regresión Logística. https://aprendeia.com/algoritmo-regresion-logistica-machine-learning-teoria/

Aqeel, M., & Rehna, T. (2020). Association among school refusal behavior, self-esteem, parental school involvement and aggression in punctual and truant school-going adolescents: a multilevel analysis. *International Journal of Human Rights in Healthcare*, *13*(5), 385–404. https://doi.org/10.1108/IJHRH-06-2020-0041

AWS. (s.f.). ¿Qué es Python? https://aws.amazon.com/es/what-is/python/

Beltrán, B. (s.f.). Definiciones de Minería de datos, Regresión Lineal, Árboles de Decisión, Redes Neuronales, Redes Bayesianas. http://bbeltran.cs.buap.mx/NotasMD.pdf

Berlanga-Silvente, V., Rubio-Hurtado, M.-J., & Vilà Baños, R. (2013). Cómo aplicar árboles de decisión en SPSS. *REIRE Revista d'Innovació I Recerca En Educació*, 6(1), 65–79. https://doi.org/10.1344/reire2013.6.1615

Boyacı, A. (2019). Exploring the factors associated with the school dropout. *International Electronic Journal of Elementary Education*, *12*(2), 145–156. https://doi.org/10.26822/iejee.2019257661

Boylan, R. L., & Renzulli, L. (2017). Routes and Reasons Out, Paths Back: The Influence of Push and Pull Reasons for Leaving School on Students' School Reengagement. *Youth and Society*, *49*(1), 46–71. https://doi.org/10.1177/0044118X14522078

Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. *Children and Youth Services Review*, *96*, 346–353. https://doi.org/10.1016/j.childyouth.2018.11.030

Constitución Política de Colombia [Const]. Art. 67. 1991. (Colombia).

Coussement, K., Phan, M., De Caigny, A., Benoit, D. F., & Raes, A. (2020). Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. *Decision Support Systems*, *135*. https://doi.org/10.1016/j.dss.2020.113325

Cunningham, P., Cord, M., Delany, SJ (2008). Aprendizaje supervisado. En: Cord, M., Cunningham, P. (eds) Técnicas de aprendizaje automático para multimedia. Tecnologías Cognitivas. Springer, Berlín, Heidelberg. https://doi.org/10.1007/978-3-540-75171-7_2

Datos Abiertos Gov.co. (2020). Resultados indicadores: deserción escolar en la ciudad de Medellín. https://www.datos.gov.co/d/nudc-7mev/visualization

De Almeida Nascimento, M. V. L., & de Andrade, M. O. (2022). School transportation program as means to improve public education in a minor rural town in Northeastern Brazil. *Ensaio*, *30*(114), 182–206. https://doi.org/10.1590/S0104-40362021002903093

Diaz, P. (2008). Modelo Conceptual Para La Deserción Estudiantil Universitaria Chilena. https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07052008000200004

Doyle, G., & Keane, E. (2019). 'Education comes second to surviving': parental perspectives on their child/ren's early school leaving in an area challenged by marginalisation. *Irish Educational Studies*, *38*(1), 71–88. https://doi.org/10.1080/03323315.2018.1512888

Escudero, J. (2005). Fracaso escolar, exclusión social. ¿De qué se excluye y cómo?, 2. https://dialnet.unirioja.es/servlet/articulo?codigo=2304464

Fonseca Grandón, G. R. (2018). Trayectorias de permanencia y abandono de estudios universitarios: una aproximación desde el currículum y otras variables predictoras. *Educación y Educadores*, *21*(2), 239–256. https://doi.org/10.5294/edu.2018.21.2.4

Gallego, M. G., Perez de los Cobos, A. P., & Gallego, J. C. G. (2021). Identifying students at risk to academic dropout in higher education. *Education Sciences*, *11*(8). https://doi.org/10.3390/educsci11080427

Gil, A. J., Antelm-Lanzat, A. M., Cacheiro-González, M. L., & Pérez-Navío, E. (2019). School dropout factors: a teacher and school manager perspective. *Educational Studies*, *45*(6), 756–770. https://doi.org/10.1080/03055698.2018.1516632

Gil, A. J., Antelm-Lanzat, A. M., Cacheiro-González, M. L., & Pérez-Navío, E. (2021). The effect of family support on student engagement: Towards the prevention of dropouts. *Psychology in the Schools*, *58*(6), 1082–1095. https://doi.org/10.1002/pits.22490

Gutiérrez-Cobo, M. J., Cabello, R., & Fernández-Berrocal, P. (2017). Inteligencia emocional, control cognitivo y estatus socioeconómico de los padres como factores protectores de la conducta agresiva en la niñez y la adolescencia. *Revista Interuniversitaria de Formación Del Profesorado*, *31*(1).

Iberdrola s.f. Definición Machine Learning, Aprendizaje supervisado y Aprendizaje no supervisado. https://www.iberdrola.com/innovacion/machine-learning-aprendizaje-automatico

IBM, (2021). Conceptos básicos de ayuda de CRISP-DM. https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

Jamaluddine, Z., Akik, C., Safadi, G., Abou Fakher, S., El-Helou, N., Moussa, S., Anid, D., & Ghattas, H. (2022). Does a school snack make a difference? An evaluation of the World Food Programme emergency school feeding programme in Lebanon among Lebanese and Syrian refugee children. *Public Health Nutrition*, 1–13. https://doi.org/10.1017/s1368980022000362

La Universidad en Internet (UNIR). (2020). Estimulación temprana: qué es y cuáles son sus ventajas. *UNIR Revista*. https://www.unir.net/educacion/revista/estimulacion-temprana/

La Universidad en Internet (UNIR). (2021). Clustering: qué es y cuál es su uso en Big Data. *UNIR Revista*. https://www.unir.net/ingenieria/revista/clustering/

Ley 1581 de 2012. Por la cual se dictan disposiciones generales para la protección de datos personales. 17 de octubre de 2012.

M. Pal. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26:1, 217-222.

Machado do Nascimento, L., & Steren dos Santos, B. (2021). Processos motivacionais de estudantes do curso de Pedagogia e suas relações para a permanência na universidade. *InterCambios*, *8*. https://doi.org/10.29156/INTER.8.1.9

Maheshwari, E., Roy, C., Pandey, M., & Rautray, S. S. (2020). Prediction of Factors Associated with the Dropout Rates of Primary to High School Students in India Using Data Mining Tools. In *Advances in Intelligent Systems and Computing* (Vol. 1013). https://doi.org/10.1007/978-981-32-9186-7_26

Makhloga, V. S., Raheja, K., Jain, R., & Bhattacharya, O. (2021). Machine learning algorithms to predict potential dropout in high school. In *Lecture Notes on Data Engineering and Communications Technologies* (Vol. 54). https://doi.org/10.1007/978-981-15-8335-3_17

Martinic Lenta, R. (2019). *What is behind higher-education permanence?* (Vol. 45).

Melgar, A. S., Garay-Argandoña, R., Aranda, E. A. E., & Hernández, R. M. (2020). Management risk factors in educational institutions and their impact on peruvian student desertion. *Elementary Education Online*, *19*(4), 226–233. https://doi.org/10.17051/ilkonline.2020.04.124

Ministerio de Educación Nacional. (s.f.). La Deserción escolar - ¿Qué es? https://www.mineducacion.gov.co/1621/articles-293659_archivo_pdf_abc.pdf

Ministerio de Educación Nacional. (2022a). Definición de deserción escolar. https://www.mineducacion.gov.co/1621/article-82745.html

Ministerio de Educación Nacional. (2022b). *Niveles de la educación básica y media*. https://www.mineducacion.gov.co/portal/Preescolar-basica-y-media/Sistema-de-educacion-basica-y-media/233834:Niveles-de-la-educacion-basica-y-media

Murillo-Zabala, A. M., & Jurado-De los Santos, P. (2020). Student permanence: Factors that influence at politecnico internacional of Bogota, Colombia. *Revista Electrónica Educare*, *25*(1), 1–25. https://doi.org/10.15359/ree.25-1.6

Noble RN, Heath N, Krause A, Rogers M. Teacher-Student Relationships and High School Drop-out: Applying a Working Alliance Framework. Can J Sch Psychol. 2021;36(3):221-234. doi:10.1177/0829573520972558

Ogresta, J., Rezo, I., Kožljan, P., Paré, M.-H., & Ajduković, M. (2020). Why Do We Drop Out? Typology of Dropping Out of High School. *Youth & Society*, *53*(6), 934–954. https://doi.org/10.1177/0044118X20918435

Observatorio de Trayectorias Educativas. 2021. Repitencia, deserción y abandono. https://ote.mineducacion.gov.co/tablero-control/repitencia.

Oracle. (s.f.). Definición de Big Data. https://www.oracle.com/co/big-data/what-is-big-data/

Pasic, D., & Kucak, D. (2020). Machine learning model for detecting high school students as candidates for drop-out from a study program. *2020 43rd International Convention on Information, Communication and Electronic Technology, MIPRO 2020 - Proceedings*, 1140–1144. https://doi.org/10.23919/MIPRO48935.2020.9245405

Pavez, A. R. (2020). Toward the prevention of school absenteeism: Proposals for socio-educational intervention. *Revista Brasileira de Educacao*, *25*. https://doi.org/10.1590/S1413-24782020250037

Procedimiento interno Alcaldía de Medellín. PR-EDUC-068 Implementación de estrategias de acceso y permanencia.

Resolución 07797 de 2015 [Ministerio de Educación Nacional]. Por medio de la cual se establece el proceso de gestión de la cobertura educativa en las Entidades Territoriales Certificadas. 3 de diciembre de 2015.

Rogelberg, S. L., Starrett, A., Irvin, M. J., & DiStefano, C. (2021). Examining motivation profiles within and across socioeconomic levels on educational outcomes. *International Journal of Educational Research*, *109*. https://doi.org/10.1016/j.ijer.2021.101846

Rubio, N. (2022). Test de Turing: qué es, cómo funciona, ventajas y limitaciones. https://psicologiaymente.com/cultura/test-turing

Santamaría, F. (2018). Estudio y predicción de activos financieros mediante redes neuronales. Universidad Politécnica Madrid. https://oa.upm.es/53810/1/TFG_FERNANDO_SANTAMARIA_VAZQUEZ.pdf

Simon, O., Nylund-Gibson, K., Gottfried, M., & Mireles-Rios, R. (2020). Elementary absenteeism over time: A latent class growth analysis predicting fifth and eighth grade outcomes. *Learning and Individual Differences*, *78*. https://doi.org/10.1016/j.lindif.2020.101822

Tasnim, N., Paul, M. K., & Sarowar Sattar, A. H. M. (2019). Performance Analysis of Different Decision Tree Based Methods for Identifying Drop out Students. *1st International Conference on Advances in Science, Engineering and Robotics Technology 2019, ICASERT 2019*. https://doi.org/10.1109/ICASERT.2019.8934518

Uliyan, D., Aljaloud, A. S., Alkhalil, A., Amer, H. S. A., Mohamed, M. A. E. A., & Alogali, A. F. M. (2021). Deep Learning Model to Predict Students Retention Using BLSTM and CRF. *IEEE Access*, *9*, 135550–135558. https://doi.org/10.1109/ACCESS.2021.3117117

Velázquez Narváez, Y., & González Medina, M. A. (2017). Factors associated with student persistence: The case of the UAMM-UAT. *Revista de La Educacion Superior*, *46*(184), 117–138. https://doi.org/10.1016/j.resu.2017.11.003

Viloria, A., Naveda, A., Palma, H., Núñez, W., & Núñez, L. (2020). Using Big Data to Determine Potential Dropouts in Higher Education. *Journal of Physics: Conference Series*, *1432*, 12077. https://doi.org/10.1088/1742-6596/1432/1/012077

Wiederhold, G., McCarthy, J. (1992). Arthur Samuel: Pioneer in Machine Learning. Published in: IBM Journal of Research and Development

Xie, K., Vongkulluksn, V. W., Lu, L., & Cheng, S. L. (2020). A person-centered approach to examining high-school students' motivation, engagement and academic performance. *Contemporary Educational Psychology*, *62*. https://doi.org/10.1016/j.cedpsych.2020.101877

14. Anexos

Anexo A. Acuerdo de Confidencialidad para Utilizar Bases de Datos de los Estudiantes de la Ciudad de Medellín

Alcaldía de Medellín

ACUERDO DE CONFIDENCIALIDAD PARA UTILIZAR BASES DE DATOS DE LOS ESTUDIANTES DE LA CIUDAD DE MEDELLIN

Licencia N° 013-2022

El Municipio de Medellín, Secretaría de Educación, en su calidad de titular de las bases de datos de los estudiantes de establecimientos educativos oficiales y no oficiales registrados en educación básica y media en el SIMAT y SIMPADE, autoriza a Gabriel Jaime Jaramillo Ciro para tomar la información en las bases de datos de SIMAT y SIMPADE, para su uso y tratamiento en el marco de su trabajo de grado, el cual está planteado para cumplir con el requisito para obtener el título de especialista en Big Data e Inteligencia de Negocios de la Universidad Católica Luis Amigó.

Siguiendo los lineamientos de la ley de Censos, Ley 79 de 1993, y el concepto de septiembre 27 de 1999 donde el Consejo de Estado, "al estudiar la reserva estadística a la luz de esta ley, establece los alcances de esta disposición señalando que no se contraen sólo a la información de la naturaleza anotada obtenida en los censos de población y vivienda, como pudiera deducirse ligeramente del epígrafe de la ley 79 de 1993", los datos suministrados por la secretaría de educación tienen un carácter estrictamente reservado, y por lo tanto no podrán darse a conocer al público sino únicamente en resúmenes numéricos, que no hagan posible deducir de ellos información alguna de carácter individual que pudiere utilizarse para fines de tributación fiscal, investigación judicial, o cualquier otro objetivo diferente del propiamente estadístico.

La información suministrada se realiza en concordancia con el artículo 13 de la Ley 1581 de 2012 en el que se establece el suministro de información a las entidades públicas o administrativas en ejercicio de sus funciones legales o por orden judicial. De igual manera los hacemos partícipes de los requisitos consagrados en el artículo 12 del Decreto Reglamentario 1377 de 2013, para el tratamiento de datos personales de niños, niñas y adolescentes:

> "El tratamiento de datos personales de niños, niñas y adolescentes está prohibido, excepto cuando se trate de datos de naturaleza pública, de conformidad con lo establecido en el artículo 7 de la ley 1581 de 2012 y cuando dicho tratamiento cumpla con los siguientes parámetros y requisitos:
> 1. Que responda y respete el interés superior de los niños, niñas y adolescentes.

42777468

Alcaldía de Medellín

2. Que se asegure el respeto de sus derechos fundamentales
(...)"

Gabriel Jaime Jaramillo Ciro se compromete para con el Municipio de Medellín, Secretaría de Educación:

1. Respetar los derechos de autor (morales y patrimoniales) que sobre la base de datos tenga el Municipio de Medellín

2. No incluir dentro del material entregado logos, símbolos, leyendas ni publicidad alguna sin autorización expresa del Municipio de Medellín

3. No realizar modificación alguna de la información mencionada en el presente documento, sin autorización expresa del Municipio de Medellín

4. Responder por la correcta utilización de dicho material

5. No ceder los derechos y obligaciones derivados del presente documento a ninguna persona natural y/o jurídica, sin previo consentimiento expreso del Municipio de Medellín, Secretaría de Educación

6. No comercializar de ninguna forma la información relacionada en el presente documento

7. Si necesita utilizar la información suministrada, para un fin diferente al relacionado en el presente documento, solicitará previamente permiso al Municipio de Medellín, Secretaría de Educación

8. Cuando requiera entregar la información a una persona natural o jurídica, para adelantar un proyecto, deberá llevar un control adecuado de los préstamos informando al Municipio de Medellín, Secretaría de Educación

9. Para cada entrega a un tercero se deberá diligenciar acta de compromiso igual a esta licencia de uso y entregarla al Municipio de Medellín, Secretaría de Educación

10. El Municipio de Medellín tendrá derecho a reclamar los perjuicios que se le ocasione por cualquier concepto en caso de no cumplir lo estipulado en esta licencia de uso

42777416

Alcaldía de Medellín

11. Una vez autorizada la licencia de uso por parte del Municipio de Medellín - Secretaría de Educación, Gabriel Jaime Jaramillo Ciro, se compromete a retroalimentar todos los subproductos generados con la información entregada.

12. Siempre que la información sea utilizada como referencia, Gabriel Jaime Jaramillo Ciro deberá sin excepción alguna referir la fuente de información del Municipio de Medellín - Secretaría de Educación

Para Constancia, se firma en la ciudad de Medellín, el día 1 de junio de 2022.

ALEJANDRA MARQUEZ MEJIA
Subsecretaría de Planeación Educativa
Secretaría de Educación
Medellín

GABRIEL JAIME JARAMILLO CIRO
Estudiante Especialización en Big Data e Inteligencia de Negocios
Universidad Católica Luis Amigó

Elaboró:	Revisó
Maryori Alzate Sánchez Cargo: Profesional Universitario Observatorio para la Calidad Educativa de Medellín - OCEM	**John Jairo Rico Valencia** Cargo: Coordinador Observatorio para la Calidad Educativa de Medellín - OCEM